UNIVERSITY OF OSLO
Department of Culture Studies
and Oriental Languages

Language Homogeneity in the Japanese Wikipedia

Master's Thesis in
East Asian Linguistics

60 Credits

Spring 2009

Karl-André Skevik

May 12, 2009

Abstract

The Internet based encyclopedia Wikipedia is a potentially very useful source of information, but intuitively, it is difficult to have confidence in the quality of an encyclopedia that anyone can modify. Several studies have been made that examine the correctness of a subset of articles, but the huge number of articles and constant changes limit the possible scope of this approach.

For an encyclopedia, one aspect of correctness is writing style, and especially for Wikipedia, an inconsistent writing style would give a bad impression; if errors that can be detected by any native speaker of a language go uncorrected, how likely is it that errors that only a subject expert can detected will be corrected?

We study the Japanese Wikipedia, because Japanese is a language where honorifics processes are very explicit, involving different forms, between which language users in some cases need to choose every time a sentence is uttered or written. Especially the difference between plain and polite forms is sufficiently easy to detect in a sentence for it to be feasible to perform this operation with a computer, allowing examination of all articles in the Japanese Wikipedia.

Using this approach, we find that the writing style of the Japanese Wikipedia is largely consistent with the style guidelines for the project. The examples of different style usage that we identified, and examined, were mainly found in articles that had only had a small number of changes made by a small number of different editors.

Acknowledgment

I would first of all like to thank my supervisor Bjarke Frellesvig. His patience has been much appreciated. I am also very grateful to Tomoko Okazaki Hansen for teaching me before I joined the master program, and for making it possible for me to travel to Japan as an exchange student.

Thanks to Reiko Abe Auestad for convincing me to finally join the master program in Japanese, and to Naomi Yabe Magnussen for providing letters of introduction allowing me to use the library at Waseda university during the spring of 2009.

Finally, thanks to Tami Aamodt for helpful feedback on my (mis)usage of linguistic terminology, and to Erik Vasaasen for proofreading the final draft of the text.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Wikipedia is a comprehensive Internet based encyclopedia that is available on the Internet in a wide range of languages. What is unusual about this encyclopedia is that the articles are largely written and maintained by volunteers, many of whom are anonymous. Anyone that wishes to participate can contribute new content and make corrections, which makes it significantly different from traditional encyclopedias.

1.1 Problem area

The permissive approach used by Wikipedia does however make it natural to question the correctness of the content, and the possibility of achieving a consistent writing style. Proving that the contents of an arbitrary article is correct might in many cases require specialized knowledge, but style inconsistencies in the text can be spotted by anyone familiar with the language an article is written in. An encyclopedia is expected to have a formal writing style consistently applied across all articles, and despite being an Internet based collaborative project, we argue that the same expectations apply to Wikipedia. An absence of style errors would not imply that an article is factually correct, but extensive style inconsistencies would make it more difficult

to trust Wikipedia articles. After all, if problems that any native speaker can detect are left uncorrected in an article, it is less likely that factual errors only correctable by subject experts will be found and removed.

This thesis examines the writing style of the Japanese Wikipedia, using an automated analysis approach in order to examine all articles. This approach is motivated by the presence in Japanese of very distinct writing styles that can serve to signal different degrees of formality and politeness. The difference in language usage is not limited only to word choice, but includes different forms for sentence-final elements. For example, verbs can generally have different forms with essentially the same meaning, but where different degrees of respect and politeness are implied, depending on context. It is generally necessary to choose between forms, with the proper form typically defined by genre. Because many of these forms are sufficiently distinguishable for computer based analysis to be feasible, it should be possible to determine the writing style of a sentence automatically. By examining all sentences in Wikipedia, we attempt to identify any inconsistencies in writing style.

1.2 Thesis overview

This thesis has the following goals: Firstly, to classify the writing style used by the majority of Wikipedia articles. Secondly, using this classification to quantify the type and extent of variations in writing style. Thirdly, by using automated computer analysis for classification, to examine the feasibility of this approach to maintaining a consistent writing style in large distributed projects like Wikipedia.

In this thesis, we make the following claims:

- The presence or absence of honorifics processes can be used to identify characteristics of language usage that are determined by genre.
- That this type of analysis can be automated and performed with a

computer.

- That this type of analysis can be used to determine whether the content in a large project like Wikipedia is written in a consistent style, and that the style can be described if it is consistent.

- That this kind of automated analysis can be used to identify cases of incorrect or inconsistent language usage in distributed Internet based projects such as Wikipedia, where anonymous users can contribute content and make modification.

As a result of the process of attempting to prove these claims, we make the following contributions:

- A theoretical analysis of honorifics processes in Japanese, with regards to the feasibility of identifying the presence of honorifics, using a computer program.

- An examination of some of the limitations of morphological analysis tools like *MeCab*, including cases where reliable detection of honorifics processes is difficult due to ambiguity that cannot easily be resolved by a computer.

- An examination of the contents of a Wikipedia page, and the steps that need to be taken to remove content that would lead to incorrect classification of the article writing style.

- A classification of aspects of the writing style actually used in articles in the Japanese Wikipedia, compared to that of the pages used by the project to hold discussions and introduce users.

- A manual examination of some cases of inconsistent language usage in the Wikipedia articles.

1.3 Structure

The thesis consists of the following chapters. Chapter 1 contains this introduction. Chapter 2 provides background information on Wikipedia, and an overview of existing research that focuses on the correctness of Wikipedia articles. A short description of some aspects of the Japanese language, with a focus on honorific processes, is given in Chapter 3. Chapter 4 discusses the amount of information required to determine the presence of the honorifics processes, with a though to the feasibility of doing this in an automated manner on arbitrary Japanese sentences. In Chapter 5, we examine the applicability of two existing systems used to classify honorifics usage in Japanese text and speech. The examination is used as a basis to determine a suited classification approach for the Wikipedia analysis.

The subject of Chapter 6 is automated analysis of Japanese sentences, with three tools for morphological analysis being compared. The results of a preliminary analysis of the Wikipedia content is described in Chapter 7, giving an overview of the expected writing style, as described by Wikipedia guidelines, in addition to the steps that are required to extract the article text from a snapshot of Wikipedia. Chapter 8 presents the results of our analysis of Wikipedia, and Chapter 9 concludes the thesis.

Chapter 2

Background and related work

This thesis studies the language usage of the Japanese Wikipedia. Background information and an overview of related work is provided in this chapter, along with a discussion of how results from related work can be expected to apply to this thesis.

2.1 Wikipedia

Wikipedia[1] (see Sanger, 2005) is an encyclopedia where basically anyone can contribute and modify content, and where the articles can be accessed and reused for free, with very few restrictions on redistribution[2]. This collaborative project has a high number of articles on a wide variety of topics and is available in many different languages. Familiarity with the following terms is useful for a discussion of Wikipedia.

Article/page. The text describing a single topic, corresponding to a single entry in a normal encyclopedia.

Editor. Generally, any user can become an *editor* by adding new content

[1] `http://www.wikipedia.org`

[2] With the restrictions primarily being a restriction on imposing additional restrictions.

or modifying the content of an existing article. Editors can receive additional administrative privileges by doing work that is perceived by other editors as being of high quality. It is possible to register and create a named account, but at the time of writing, being a registered user is not necessary to make changes to most Wikipedia articles.

Change history. An editor submits a change to an article along with a comment describing the change. A history of changes is maintained, allowing past versions to easily be restored in case of vandalism. Because information identifying the editor is included in the change history[3], it can also be used as an aid in identifying vandals.

Discussion pages. Any article can have a discussion page, which can be used by editors to handle disagreements, discuss needed changes, etc.

User pages. Personal pages where editors can introduce themselves and describe their Wikipedia activities.

Featured page. The articles that have been referenced on the Wikipedia front page are referred to as featured pages. They are considered to be the articles in Wikipedia with highest quality. Articles must pass through a peer-review and quality assurance process to qualify for this status, which can be lost if the quality is not maintained over time.

Vandalism/vandals. Users that delete content or make malicious modifications are referred to as vandals. With a large number pages and a high profile, being able to easily detect and correct acts of vandalism is necessary in order for Wikipedia to maintain article quality.

There are however many potential problems that can affect an encyclopedia that anyone can modify. Denning et al. (2005) list some problems related

[3]The Internet address of the machine used to submit the change is stored for unregistered users. Registered users are identified by their username.

to relying on the content in Wikipedia, including volatility caused by changes over time, and uncertainty that the content is accurate and complete. This is not only a theoretical problem; incorrectness can result not only from lack of accurate knowledge or accidental mistakes made by article writers, but from malicious users that deliberately introduce incorrect information, spam, or make other similar undesired changes (see Lorenzen, 2006).

However, if all articles and corrections had to be peer-reviewed by professionals before being published, it is likely that Wikipedia would not have reached the size and popularity it has today. The openness of Wikipedia has clearly succeeded in producing a large amount of content, and it is possible to argue that it is sufficiently popular for mistakes and errors to be quickly noticed and corrected. Even though Wikipedia might never become suitable for citation in academic work (see Waters, 2007), it can still be very useful to many people if the content can be regarded as reliable. The actual correctness of Wikipedia has for this reason been studied by several researchers.

2.2 Related work

Most of the existing research falls into one of four categories. The first category includes research that looks at the procedure for avoiding incorrectness and detecting errors, the second, research that attempts to evaluate the correctness of the actual content. A variation of the second category is research that looks at the completeness of existing content, while the final category consists of work that proposes ways in which the quality of content can be automatically classified or maintained.

2.2.1 Process evaluation

Lorenzen (2006) looks at one technique used by Wikipedia editors for vandalism detection. Over a period of several months, the author examined the contents of a page used to report potential vandalism and found that a

significant amount of resources are spent on addressing this issue, including vandal detection, correction of vandalized pages, and potentially the banning of users responsible for repeated acts of vandalism. Controversial and frequently vandalized pages are likely to be quickly corrected, but as the author concludes, detecting this in less popular articles can be difficult, especially if the vandalizing user is careful.

Viégas et al. (2004) have tried to determine how a system as open as Wikipedia can actually work. As the history of changes for Wikipedia articles are publicly viewable, this information was used to visualize the changes that occur for a page over time. Acts of vandalism were found by the authors to have been corrected fairly fast, with a median time of 1.7 minutes. Editors can request a notification each time specific pages are modified, which is one possible reason for the generally quick response to undesired changes. Disagreements between editors can however result in so-called *edit wars*, where a contested piece of text can be repeatedly changed back and forth. In a follow-up work, Viégas et al. (2007a) examines collaboration mechanisms two years later. By looking at the article discussion pages and manually classifying entries in these pages, the authors observe an increased degree of coordination and planning among Wikipedia contributors. An even more recent work by Viégas et al. (2007b) finds that for some types of articles, such as the featured articles, Wikipedia has developed an elaborate editing and peer-review process.

Stvilia et al. (2008) study several issues related to the perception of article quality in the Wikipedia community, and the processes involved in quality assurance. Based on the contents of the discussion pages for 60 articles, the authors find that the English Wikipedia community has developed extensive processes to achieve article quality, including criteria for quality assessment and mechanisms for giving editors that do good work additional privileges. Instead of error prevention, Wikipedia makes use of techniques that allow problems to be fixed quickly when they occur (see Stvilia et al., 2008, p. 33).

2.2.2 Correctness

Emigh and Herring (2005) has looked at the extent to which Wikipedia, and one other similar project, produce work that is similar or dissimilar to existing print encyclopedias, in other words, whether they belongs to the same genre. This was done by looking at the formality of the language in use. Based on the content of 15 articles, the degree of formality was quantified by separately counting the occurrence of word usage typical of both formal and informal English language usage genres. For the informal genres this included contractions and personal pronouns, while noun formative suffixes was used to measure the degree of formality. In addition, the average word length and number of words in a sentence were calculated. The results for these articles show that the level of formality is close to that of the print encyclopedia, while the content of the discussion pages is far more informal.

Nielsen (2007) examined the use of citations of scientific journals in the English Wikipedia, and finds these to be used in an increasingly structured manner, having citation usage correlating with that of scientific journals.

An expert-led peer-review performed by Nature (see Giles, 2005) compared Wikipedia to Encyclopedia Britannica, by examining 42 articles. Errors were found in both encyclopedias; with an average of three errors found in the science articles of the Encyclopedia Britannica, compared to an average of four in Wikipedia. Rector (2008) has done a similar comparison, looking at nine randomly selected history related entries in Wikipedia and three other sources, including two reputed subject specific encyclopedias. The author identified a larger number of incorrect and unattributed facts in the Wikipedia entries, but errors were also found in the other sources.

A survey of the credibility of Wikipedia has been performed by Chesney (2007), by asking academics to rate the credibility of two articles, one related to their field and one randomly selected article. The participants found the articles in their own field to be more credible than the randomly selected articles, possibly indicating that Wikipedia is perceived as less credible than

it actually is. However, errors were identified in 13% of the articles.

Wilkinson and Huberman (2007) have looked at the correlation between article quality and factors such as the number of edits and distinct editors for an article. The set of featured articles, regarded as being of high quality, was used as reference. The authors determined that there is a strong correlation, with a high number of editors and edits being indicative of high article quality. However, most of the articles have relatively few edits compared to the high-quality articles. The number of edits and editors of an article is also used by Lih (2004), with a focus on articles that have been cited by newspapers, magazines, and similar news sources. The results indicate that increased attention leads to improved article quality. A similar approach to automatically determine article quality is used by Blumenstock (2008), with word count as the quality indicator.

Luyt et al. (2008) have studied articles that have contained errors, looking at when the errors were introduced and how much time passed before they were fixed. Many errors were found to have been added early in the life of the article, and for almost one fifth of the articles, in the first version of the article. Many of the later article changes modified the language used in the article rather than the content (see Luyt et al., 2008, p. 328).

The quality of contributions to the French and Dutch versions of Wikipedia by different types of users has been studied by Anthony et al. (2007). They observed that the highest quality contributions, measured in the extent to which these contributions had be been retained over time, had not been contributed by registered users, but by unregistered users with few changes made in total. The authors speculate that unregistered users that make only a small number of additions or changes represent specialists that make contributions in a single field, or readers that notice and correct minor errors or missing data. A similar study has been done by Kittur et al. (2007), who has looked at modifications to the English version of Wikipedia. The percentage of contributions made by administrators and frequently contributing

users was found to have decreased during the lifetime of Wikipedia, as the largest growth has been in the number of users that have made less than 100 changes. However, their findings suggest that the largest text contributions are made by administrators and the group of most active users, while the least active users make changes that overall reduce the number of words in an article. The difference in contribution levels is confirmed by Ortega et al. (2008). Looking at the number of contributions by registered users for the ten largest language versions of Wikipedia, they observe that the majority of changes are made by a small group of users when there are few authors and articles. However, as the number of editors and articles increases, the changes are distributed more equally among the users. The exception is the Japanese version of Wikipedia, which compared to the other languages have the number of changes distributed among a relatively high number of editors, relative to the number of articles and registered users. The authors do not examine the reason for the difference, but possibly it is related to a fact observed by Voss (2005), that the Japanese Wikipedia has a relatively high number of changes made by unregistered users, which are not included in the results obtained by Ortega et al. (2008).

2.2.3 Completeness

Devgan et al. (2007) study the accuracy of medical information in Wikipedia, using two independent reviewers to examine a selection of articles on common medical procedures. The authors found that, though not complete, the Wikipedia entries were accurate.

Luyt et al. (2007) compare the entries related to Biochemistry to those found in the Encyclopedia Britannica, with a first year university textbook on the subject used to identify concepts that ought to appear in both encyclopedias. Wikipedia was found to be more comprehensive, but both encyclopedias were far from containing all the concepts described in the textbook.

The completeness of drug information in Wikipedia is studied by Clauson

et al. (2008), by comparing it to a specialized drug database. A set of questions regarding medical drug information was constructed and independently verified. The extent to which the specialized drug database and Wikipedia was able to answer these question was then verified, and though not factually incorrect, Wikipedia was found to be less complete. A similar study for medical informatics done by Altmann (2005) also found many basic concepts to be missing.

2.2.4 Automated process proposals

HU et al. (2007) propose a way of automatically calculating article quality rankings based on the retention of content changes and additions made by editors. Article quality is calculated based on the authority of the article editors, with editor authority depending on the quality of the articles to which an editor has contributed. Potthast et al. (2008) suggest a way of automatically detecting vandalism by looking at characteristics of typical changes made by vandals.

Adler and de Alfaro (2007) propose a reputation system for Wikipedia editors, based on the degree to which changes made by editors remain in Wikipedia. Similar approaches for automatically analyzing the quality and trustworthiness of article content is proposed by Dondio and Barrett (2007), and McGuinness et al. (2006).

2.3 Discussion

The relevance of the problem of assessing reliability is in other words well understood, with a wide variety of approaches used to analyze it. The work closest to this thesis is that of Emigh and Herring (2005), which compares the genre of Wikipedia to that of traditional print encyclopedias. This thesis makes a similar survey of the Japanese Wikipedia, looking at the formality

and consistency of language usage, but rather than manually examining a small number of articles, we study the entire Japanese Wikipedia.

The results of the related work listed above generally indicate that the content of Wikipedia is generally of high quality and comparable to that of traditionally produced works such as the Encyclopedia Britannica. Errors might exist, but will usually be fixed quickly, especially in articles that receive a lot of attention. There are however many pages, most of which receive few changes, and errors in these pages are more likely to go undetected.

Based on the above examination of related work it is to be expected that improper language usage in heavily edited articles should be quickly fixed. Barring use of automatic language analysis by Wikipedia editors, it should primarily exist in articles with few edits and editors.

2.4 Wikipedia citation

The potentially high rate at which content in Wikipedia can change causes a citation problem. Referring to a specific article is not reliable because the content can undergo significant change at any time. However, the change history system used by Wikipedia maintains a copy off all versions of a page[4], making it possible to reference a single version of a given article. Even if the article is subsequently changed, the cited content will still remain the same as long as the correct version is accessed. For this reason we specify the last modification date whenever we in this thesis refer to the contents of an article.

[4]It is possible for versions to be deleted (see Stvilia et al., 2008, p. 12), but this will likely not be an issue for articles that do not contain spam or similarly problematic content.

2.5 Summary

This chapter gives an overview of related work and identifies several characteristics of Wikipedia. Chapter 8 includes a comparison of our results with those of the related work examined in this chapter.

Chapter 3

Japanese language

This chapter gives an overview of the Japanese honorifics system, which in Japanese is called *keigo* and includes a wide range of honorifics processes. The focus in this chapter is on elements that are relevant for identifying genre differences, including sentence structure.

3.1 Japanese honorifics

Wetzel (2004) gives an overview of some of the research that has been done by both Japanese and Western scholars in attempting to analyze honorifics in the Japanese language. One useful definition is that of the separation between types of expression alternatives with the same meaning, and the factors that govern the process of choosing between them (see Wetzel, 2004, p. 39). Expression alternatives include politeness, roughness, formality, elegance, and vulgarity. The deciding factors include location, whether the context is written or spoken, interpersonal relationships, and psychological factors such as the intent of the speaker, or the extent to which the speaker understands or is aware of the context. In this thesis, the context is the Japanese version of Wikipedia (what can be considered the correct writing style is examined in Section 7.1). In this chapter, we look at the range of

possible expression alternatives related to honorifics.

One way of defining the expression alternatives in Japanese is through the three primary honorific processes in the language. Given a speaker, an addressee, and a referent, where the referent can be the speaker, the addressee, or a third party, Shibatani (1991, p. 375-376) describes Japanese as having honorific processes along two independent axes: the speaker-addressee axis and the speaker-referent axis. The first of these is also referred to as *addressee controlled honorifics*, while the speaker-referent axis consists of so-called *subject honorifics* and *object honorifics*. These three terms roughly correspond to the Japanese terms *teineigo* (polite language), *sonkeigo* (respect language), and *kenjougo* (humility language). A fourth related category, listed in Wetzel (2004, p. 29), is *bikago* (beautification language).

3.1.1 *Teineigo*

According to Matthews (2007), the term addressee controlled honorifics generally refers to language used by a speaker to show deference to the addressee. This is also the case with *teineigo*, which "indicates an attitude of respect on the part of the speaker for the hearer" (see Wetzel, 2004, p. 30). The so-called plain form is the alternative to the polite form; Cook (1998, p. 1) describes it as the *non-honorific counterpart* to the polite form. It is generally used in informal situations, even though, as we see below, this view is somewhat simplistic.

Polite verb forms have the *-masu/-mashita* verbal endings attached to the verbal stem (see Shibatani, 1991, p. 375), as can be seen in the example below, taken from Wetzel (2004, p. 5). The first part of the verb is the same in both cases, but compared to the plain variant in (1a), the polite variant in (1b) is longer and has the *-masu* ending.

(1) a. Plain form
shiru
"to know, find out"

b. Polite form
shirimasu
"to know, find out"

For the copula, the plain variant uses the *da/datta* forms, while the polite variant uses the *desu/deshita* forms (see Shibatani, 1991, p. 375). In (2), taken from Wetzel (2004, p. 5), the two *da/desu* variants of the copula are shown. The plain form can additionally have ellipsis of the copula (see Kaiser et al., 2001, p. 96).

(2) a. Plain form
Tanaka-san da.
"I/he/she am/is Tanaka."
b. Polite form
Tanaka-san desu.
"I/he/she am/is Tanaka." (polite)

Informal spoken Japanese will generally use plain forms, along with the many language variations typically found in colloquial speech. Examples include particles such as *ne* and *yo* and extensive ellipsis (see Shibatani, 1991, p. 360). Formal spoken Japanese will generally use the polite forms, and possibly the other honorific processes described below. Mixed usage, one speaker using polite forms and another replying with plain forms, is generally accepted to be indicative of different social status between the two speakers, but Cook (1998) argues that this view of the the two forms is too simple. Native speakers will often shift between the two without this necessarily being connected to social status. For example, the use of the plain form in a neutral tone, without any final emotional particles, might be used when the focus is on the information in the sentence. The use of the plain form will then, according to Cook (1998, p. 98), not be a reference to relative social status or formality, and will not sound rude.

A similar duality in usage of the plain form can be seen in written Japanese, which will generally be quite formal, to the extent that, for letters, polite forms are likely used even between members of the same family, according to Shibatani (1991, p. 360) and Musteric (2003). In letter writing, the polite form is also often used in conjunctions, and when modifying nouns (see Tatematsu et al., 1997, p. 20). However, Shibatani (1991, p. 360) notes that even if polite forms are used, along with a high degree of formality, in writing with a known recipient, this is not typical when there is no specific reader. Ellipsis of the copula, as shown in (3), taken from Kaiser et al. (2001, p. 97), is not uncommon in newspapers and other writing styles. Particles like *ne* and *yo* are generally not used in writing (see Shibatani, 1991, p. 360).

(3) Copula ellipsis *arashi no ato no shizukana asa ().*
"A quiet morning after the storm."

Furthermore, ellipsis of predictable verbs is not uncommon in newspapers (see Makino and Tsutsui, 2002b, p. 41), along with use of the *shi* connective form of the verb *suru* (to do). Newspapers and scholarly articles also use the plain form and the *de aru* variant of the copula. Use of *de aru* is not common in spoken language, and the polite form, *de arimasu*, is primarily found in speeches and formal business letters (see Makino and Tsutsui, 2008, p. 35). A study of *de aru* in scientific articles has been done by Lucas (1991), who observed *de aru* being used to a much larger degree than *desu* or *da*. Generally found at the end of a sentence and followed by full stop, many *de aru* sentences follow the form *X wa* NOUN *dearu*. Rather than adding information, the copula form can have an assertive function related to the theme of a sentence.

Overall, it can be said that the choice between the plain and polite form is not simply a factor of situational formality but is highly genre-dependent.

3.1.2 *Sonkeigo*

Subject honorifics indicate respect for the subject of a sentence (see Matthews, 2007, p. 30). In Japanese, the name for the subject honorifics processes is *sonkeigo* (respect language) and can involve several mechanisms (Shibatani, 1991, p. 283):

Firstly, use of an indirect reference to the actions of the subject, with the construct *o-verb ni naru*. This is shown in (4), taken from Shibatani (1991, p. 283). The addition of *ni naru* results in the actions of the subject being described as non-volitional, ascribing a similar amount of respect to the subject as would be given to a natural phenomenon (see Ivana and Sakai, 2007, p. 186).

(4) a. Plain form
Kakehi sensei ga waratta.
"Professor Kakehi laughed."

b. Subject honorific form
Kakehi sensei ga o-warai ni natta.
"Professor Kakehi laughed." (honorific)

While there might be some disagreement among linguists about the role of the elements in this construction, Ivana and Sakai (2007, p. 181) argue that the *o-* prefix carries the honorific meaning. This is consistent with the honorific function that the prefix has when used in front of nouns or adjectives (see below), because the verb in the construction is in the noun-like adverbial form (see Shibatani, 1991, p. 218).

Secondly, when, for example, referencing objects that belong to a respected person, the honorific *o-* prefix is used before the noun. This usage is again related to *bikago*, discussed in Section 3.1.4. For words of Chinese origin, the prefix will generally be *go-* rather than *o-* (Makino and Tsutsui, 2002a, p. 346).

Thirdly, through the use of the *-rare* suffix (see Shibatani, 1991, p. 375). This suffix is homophonous with the suffix for the passive, potential, and spontaneous forms, as shown in (5) from Oshima (2008). The sentence composition is slightly different, but the verb is identical in both sentences despite the meaning not being the same.

(5) a. Passive form
Taro ga shikarareta.
"Taro was scolded."

b. Subject honorific form
Sensei ga Taro wo shikarareta.
"The teacher scolded Taro." (honorific)

Fourthly, via suppletive forms, which exist for many verbs. A typical example, taken from Wetzel (2004, p. 4), is the subject honorific form of the verb *shiru* (to know/find out), which is *gonzonji (da)*). Around thirty suppletive forms exist for the more common verbs, and these need not be of the same type; *shiru* is a verb while *gozonji* is a noun.

3.1.3 *Kenjougo*

Shibatani (1991, p. 375) classifies *kenjougo* as corresponding to object honorifics, which according to Matthews (2007) indicate respect for the object of a sentence. As with *sonkeigo*, there are several mechanisms that fall into the category of *kenjougo*.

Firstly, the generic construct *o-verb suru* (see Shibatani, 1991, p. 375), which is very similar to the *o-verb ni naru* subject honorific construct discussed above. While the subject honorific process indicates respect for the subject by describing the actions of the subject as non-volitional, the object honorific construct uses the verb *suru* (to do), which is volitional. Because this requires insight into the intentions of the subject, this implies closeness towards the subject, and therefore lack of respect. As a result of this, the

respect is directed towards the object of the sentence (Ivana and Sakai, 2007, p. 186). An example of this object honorific form is shown in (6), taken from Shibatani (1991, p. 376).

(6) a. Plain form
Tarou ga sensei wo tasuketa.
"Taro assisted the teacher."

b. Plain, object honorific form
Tarou ga sensei wo o-tasuke-shita.
"Taro assisted the teacher." (humble)

Secondly, through suppletive forms (these are different from the subject honorific suppletive forms). For example, the object honorific form of the verb *shiru* (to know/find out) is *zonjiru* (see Wetzel, 2004, p. 4). The object honorific form is different from the subject honorific form (*gozonji (da)*), and both of these are different from the plain form of the verb.

Thirdly, the use of the honorific prefix *o-* with nouns or adjectives related to the object, in the same way as described above for subject honorifics (see Shibatani, 1991, p. 374).

3.1.4 *Bikago*

As noted above, the honorific *o-* prefix can be used with nouns as part of the subject and object honorification processes, typically when referring to items belonging to a person the speaker wishes to show respect towards. The usage of the so-called beautification language is identical, in that the *o-* prefix can attach to items, but rather than items belonging to a respected person, they can attach to items belonging to the speaker. It can be debated whether *bigago* should be classified as part of *keigo* or not (see Wetzel, 2004, p. 38), but it is clearly related.

The purpose of the *o-* prefix in *bigago* is not honorification, but beautification (see Shibatani, 1991, p. 374), and to demonstrate the quality of the

language used by the speaker (Wetzel, 2004, p. 4). This usage of the *o-* prefix is more common in the speech of women than in the speech of men, to the extent that over-usage is sometimes considered a problem (see Wetzel, 2004, p. 117).

In addition to this deliberate addition of the *o-* prefix, there are some examples of Japanese words where the prefix is generally always used, such as *o-sake* (alcoholic drink) and *go-han* (food/meal) (see Kaiser et al., 2001, p. 189).

3.1.5 Pronominals and sex dependent speech

Differences in the speech of men and women are not limited to the use of the *o-* prefix. This is also the case with pronominal forms (see Shibatani, 1991, p. 371), of which proper usage is determined not only by the level of formality, but also often by the sex of the speaker. Alternatives spans from the formal first person *watakushi* (gender neutral) to the very informal *ore* (male) or *atashi* (female).

Sex dependent language differences can also be observed in relation to sentence-final particles (see Shibatani 1991, p. 373, and Mcgloin 1990, p. 24), with exclamatory particles such as *wa* primarily found in female speech, and *ze* and *zo* in male speech.

3.1.6 Other factors

From the discussion above it can be seen that there are many elements that can effect formality and politeness. The overview given in this chapter is far from complete, but it covers the primary honorific processes. Wetzel (2004, p. 34) provides a list of some additional relevant elements, including so-called *minus* keigo, viz. abusive language, arrogant expressions, etc. Other important factors are vocabulary, with different vocabulary appropriate for spoken and written language, and compositional aspects such as sentence

length.

For spoken language there are additional non-linguistic elements that are relevant for politeness, such as attitude and manner (see Wetzel, 2004, p. 34), but these factors are less relevant for written language[1].

3.2 Sentence structure

Above, we describe several honorific sentence elements, but to know where they can typically be found in a sentence, it is useful to have an basic understanding of the Japanese sentence structure. The language permits some reordering of sentence elements, but is basically a Subject Object Verb (SOV) language that requires the verbal element to come last (see Shibatani, 1991, p. 259). However, non-verbal elements can still occur at the end of a sentence, especially in colloquial speech. Shibatani (1991) describes this as being caused by the non-verbal elements being added as an *afterthought*, with intonation indicating that the verbal-element is still considered to be sentence-final. Another reason can be to emphasize what would normally be the sentence-final part (see Kaiser et al., 2001, p. 197).

A minimal Japanese sentence consists of a single predicate with zero or more noun phrases. Ellipsis of the noun phrases is possible. The predicate can be followed by one or more particles, and can be a verb, an adjective, or a noun or adjectival noun with the copula. The noun phrase can consist of a noun (along with one or more particles) or an adverbial element (see Kaiser et al., 2001, p. 441, and Makino and Tsutsui, 2002b, p. 55). Some examples of minimal sentences are given in (7).

(7) a. Imperative verb
yame.
"Stop!"

[1]Less relevant, but not absent; the type of paper, surrounding illustrations, etc. are possible non-linguistic elements that are relevant for written text.

b. Adjective with sentence-final particle
takai yo.
"(Something) is expensive."

c. Noun with copula
kuruma datta.
"It was a car.'

More complicated sentences can have more than one verb, but generally it is only for the sentence-final verb that the choice between plain and polite forms exist. For example, a verb modifying a noun will generally be in the plain form (see Kaiser et al., 2001, p. 566) (but can be in the *-masu* form), and in the case of conjunctions a verb will generally be in the conjunctive form (but can also here use the longer *-masu* form (see Kaiser et al., 2001, p. 82)). Additionally, a sentence can contain a full quoted sentence, potentially with polite forms, inside quotation marks in the sentence (see Kaiser et al., 2001, p. 446).

3.3 Summary

This chapter gives an overview of Japanese honorific processes and sentence structure. An analysis of how the honorific processes described in this chapter apply to the problem area of this thesis is presented in Chapter 4.

Chapter 4

Honorifics analysis

This chapter examines what knowledge is required to analyze the various honorifics processes described in Chapter 3, in order to identify Japanese genre characteristics related to language usage. The purpose is to determine the feasibility of doing a similar analysis with a computer.

4.1 Honorific process types

We start by examining the honorifics categories at a high level, then study the various honorifics processes in more detail in Section 4.2. As noted in Chapter 3, Shibatani (1991, p. 375) describes Japanese as having honorific processes along two independent axes. Figure 4.1 illustrates this relationship, with the vertical axis representing addressee controlled honorifics (*teineigo*) and the horizontal axis representing subject and object honorifics (*sonkeigo* and *kenjougo*).

Based on what is the proper honorifics usage in different genre types, it would be possible to place the different spoken and written genres of Japanese into one of the four quadrants in the figure, resulting in the following four categories, each describing honorifics related characteristics of the genres in the category.

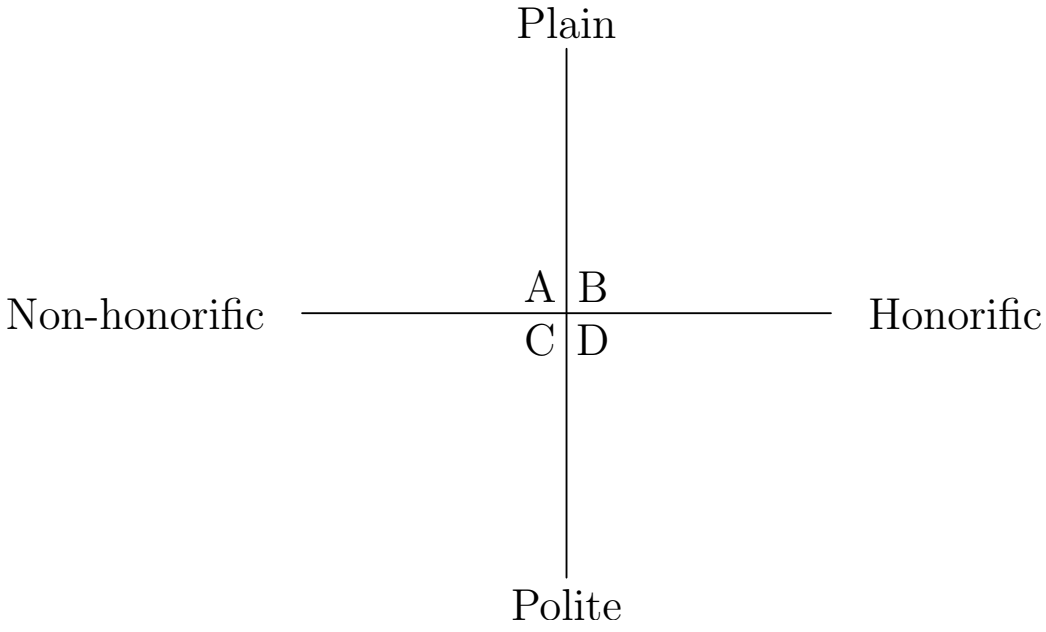


Figure 4.1: Honorifics axes

Quadrant A. Plain, no honorifics.

Quadrant B. Plain, with honorifics.

Quadrant C. Polite, no honorifics.

Quadrant D. Polite, with honorifics.

The A quadrant corresponds to colloquial speech, having plain verb and copula forms, but no usage of honorifics, even when this would be appropriate. The B quadrant includes expressions with plain verb forms, which make use of honorifics when appropriate. This combination will be the norm for informal spoken language because honorifics are generally used when talking about a respected person, even in an informal setting (see Shibatani, 1991, p. 377). In the C quadrant, can be classified formal expressions that do not make use of honorifics, even when talking about a respected person. This is according to Shibatani (1991, p. 377) not common, but possible, such as

when employees in a company talk about the company president to someone outside the company (see Shibatani, 1991, p. 379). The final quadrant corresponds to the norm for conversations with, or written correspondence to, a respected person, having polite language and proper use of honorifics.

This type of classification can be used to describe honorifics related genre characteristics. As an aid in identifying the genre to which a given text belongs, it would be useful to be able to categorize an arbitrary text based on honorifics usage. However, there are some practical problems which would complicate this task. To properly analyze these issues it is again useful to start with the separation noted in Wetzel (see 2004, p. 39), between the alternatives for making an expression, on the one hand, and the factors (such as context and interpersonal relationships) that influence their choice, on the other.

To start with the first part, to meaningfully be able to classify an expression as belonging to any of the categories requires the existence of alternatives that would be classified differently. On a purely syntactic level, this is not difficult to satisfy. From Chapter 3 it can be seen that for *teineigo*, both verbs and the copula can use either a plain or polite form. Ellipsis represents a third alternative related to the plain form. For *sonkeigo* and *kenjougo*, the same is the case; verbs, nouns, and adjectives can be made honorific, or left as non-honorific. What remains would be expressions that lack verbs, nouns, and adjectives, which would make communication of any significance difficult. As described in Section 3.2, a sentence needs a single predicate, which requires at least one of these elements.

Unfortunately, classification on a purely syntactic level is not unproblematic. For example, take (6a), on page 21. The sentence contains a verb, for which an object honorific form can be constructed (*o-tasuke shita*), giving (6b). Because both sentences use plain verb forms, a classification based only on semantics would place (6a) in quadrant A, and (6b) in quadrant B. So far, this is unproblematic, but doing this type of classification requires

knowledge beyond what can be determined from syntax validity. The object in this sentence is *the teacher*, a person worthy of respect, which makes the object honorific version in (6b) appropriate (depending on context), but if only syntax is considered it would be possible to construct the sentences in (8).

(8) a. Plain form
**Tarou ga o-sensei wo tasuketa.*
"Taro assisted the teacher."

b. Plain, subject honorific form
**Tarou ga sensei wo o-tasuke ni natta.*
"Taro assisted the teacher." (humble towards Tarou)

In (8a), the fact that *sensei* (teacher) is a noun is used along with the addition of the honorific prefix *o-* to create a sentence which is valid if the grammatical rules are applied in a very simple and mechanical fashion. The sentence in (8b) is similar, in that it uses the *sonkeigo* rules for humble verb form creation to construct a syntactically valid sentence, but one which implies respect towards *Tarou* rather than the teacher. This might be intended, but the lack of any title after the name makes the sentence sound odd without any explaining context. A human, especially a native speaker, is likely to be able to quickly determine validity or invalidity of these kinds of alternatives, but ensuring that a computer is able to do the same thing quickly becomes non-trivial. The usability of the *o-* prefix with nouns can clearly not be determined only by considering a noun in isolation from the sentence it occurs in. It becomes necessary to analyze and understand the structure of the sentence in order to identify instances where the *o-* prefix could be used correctly. As long as this can be done based on isolated analysis of the structure of a single sentence it might feasible, otherwise the problem quickly becomes difficult.

For example, what information is necessary to determine that the subject honorific form (8b) is probably incorrect, while the object honorific version of

the same sentence is acceptable? There are two persons in the sentence that can be referred to respectfully, *Tarou* (a name) and *sensei* (teacher). In this case, the fact that *Tarou* is referred to in a very familiar way, while *sensei* is a title to which respect is usually shown, might be sufficient. In other cases, such as (9), where ellipsis of the subject makes a similar comparison difficult, it might be necessary to obtain this information from other parts of the text.

(9) Plain form
saigo ni sensei wo tasuketa.
"(He/She) assisted the teacher last."

For a computer to do this would require not only syntactic parsing of the Japanese text, but semantic understanding. Moreover, there is no guarantee that the required information is found in the text. A writer might not explicitly state what anyone in the target audience is assumed to know.

Based on this discussion we can define four levels of context within which an analysis operation can be performed, shown in Figure 4.2. For each level the amount of knowledge required increases, and with it the complexity of doing analysis with a computer. One the first level, words (including lexemes and morphemes) can be analyzed in isolation, and in some cases, such as for identification of *teineigo* forms and pronominals usage, this will be sufficient. On the second level, understanding of the structure of a full sentence is required, on the third level, the contents of the full text must be understood. On the fourth level, context dependent information outside the text must also be available.

To summarize, in order to categorize a sentence as belonging to one of the four quadrants in Figure 4.1, it is necessary to be able to identify the cases where use of an honorific form would be syntactically and semantically correct, but it is not used. Doing this reliably would however, in the worst case, require information about the context within which the sentence is made or knowledge about proper *keigo* usage. While encoding this information in a computer program might be possible, it would be outside the scope of this

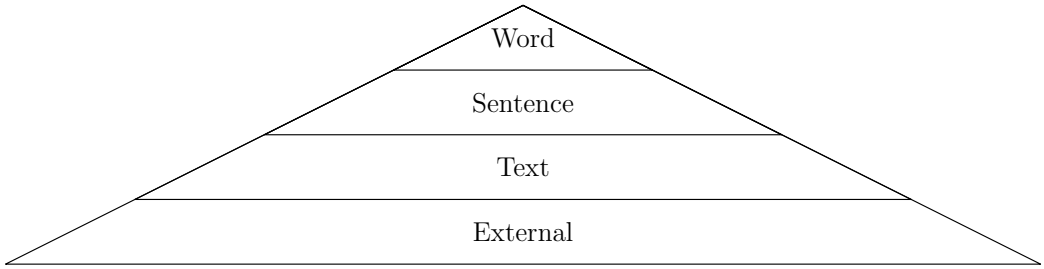


Figure 4.2: Context levels and required knowledge

thesis. The consequence of this is that for a sentence with no honorifics, we cannot reliably determine if the lack of honorifics is because usage would be semantically incorrect, or because the non-use of honorifics was deliberately chosen. What can be determined is whether plain or polite forms are used, and whether honorific forms are present in a sentence. The absence of honorific forms in a sentence does however not imply that it should be classified as being in quadrant A or quadrant C, because classification on that level of detail unfortunately appears to be difficult to reliably do with a computer.

4.2 Honorific process analysis

After having considered the possibility of honorifics analysis on a high level, we now examine in more detail the possibility of doing this for the honorifics processes from Chapter 3.

4.2.1 Polite and plain *teineigo* forms

The *teineigo* forms are relatively easy to categorize. Except in cases of ellipsis, both verbs and the copula need to use either plain or polite forms. This choice will always exist for the sentence-final verb or copula, but it can also sometimes be found inside sentences, such as in formal letters that make use of polite *-masu* forms in conjunctions. To determine if polite forms are used, the sentence level context must be considered, but primarily to identify the

final word. The preceding parts of a sentence can be examined to identify additional genre characteristics, such as the type of conjunctive form in use.

One potential problem is ambiguity caused by ellipsis, which can make correct classification more difficult. An example of a kind of ellipsis of the sentence-final verb typically found in newspaper articles is shown in (10b), taken from Makino and Tsutsui (2002b, p. 41).

(10) a. Full sentence
duponjyapan ga hatsu no nihonjinsyatyou wo ninmei suru.
"Du Pont Japan appoints first Japanese president."

b. Newspaper variant (with ellipsis)
duponjyapan hatsu no nihonjinsyatyou.
"Du Pont Japan appoints first Japanese president."

The problem in this case is that ellipsis of the sentence-final copula in (11a), a sentence with a different meaning, would result in (11b), a sentence identical to (10b).

(11) a. Alternative sentence
duponjyapan hatsu no nihonjinsyatyou da.
"(I am/he is/she is) Du Pont Japans first Japanese president."

b. Alternative sentence (with ellipsis)
duponjyapan hatsu no nihonjinsyatyou ().
"(I am/he is/she is) Du Pont Japans first Japanese president."

If analyzed in the context of the whole article, it would likely be obvious that (10a) is the correct full version of (10b), but this cannot reliably be determined on the sentence level.

4.2.2 The *o-verb ni naru sonkeigo* construct

To identify this construct it is necessary to perform the analysis on a sentence level, because it consists of multiple elements. The process is made easier

because the verb will generally be at the end of the sentence, and all parts of the construct are connected. The only variable part is the verb, but it will always be in the adverbial form.

A potential problem is that some adverbial verb forms have become normal verbs (see Shibatani, 1991, p. 218). If the classification is performed with a computer, and any of these verbs are used in this construct, they might be classified as nouns rather than verbs in the adverbial form, causing the first part to be incorrectly classified as *o-noun* (a noun with the honorific *o-* prefix). An example of this type of problem is shown in Section 4.2.3 below, for the *o-verb suru* construct.

One possible, if somewhat awkward, approach would be to let nouns that have developed from adverbial verb forms, be interpreted as adverbial verb forms, rather than nouns, when followed by *ni naru*.

4.2.3 The *o-verb suru kenjougo* construct

From an analysis point of view, this construct is not significantly different from the *o-verb ni naru* construct; multiple elements at the last part of the sentence must be examined in order to identify it.

There is a theoretical possibility that a noun identical to the adverbial verb form exists in common usage. One such example, given by Shibatani (1991, p. 218), is the noun *tasuke* (help), which is identical to the adverbial verb form of *tasukeru* (to help). This verb occurs in the object honorific expression in (6b), on page 21, as *o-tasuke shita*, and the adverbial form is identical to the noun *tasuke*. If only the noun and the *o-* prefix are considered, it would be incorrectly classified as a noun preceded by the *o-* prefix. One possible solution would be to assume that the adverbial form is the correct interpretation if a noun with the *o-* prefix is followed by the verb *suru*.

4.2.4 The *o-noun/adjective* forms

Identifying this construct is fairly simple, it merely consists of the *o-* prefix in front of a noun or adjective. As noted above, some adverbial verb forms exist as nouns in Japanese, creating the possibility that the *o-* prefix followed by a verb in the adverbial form can be incorrectly identified as a noun. However, even if we ignore this possibility, there are difficulties with this construct due to its productivity.

The *o-* prefix can occur before nouns as part of *sonkeigo* or *kenjougo* honorification, *bikago* beautification, or as part of normal noun usage in words such as *o-sake* (alcoholic drink). Handling the last set of words should be unproblematic, because they are known and can be listed as exceptions, but separating honorification and beautification is more difficult. Take (12), without any context, it is impossible to know whether the prefix is used for the purpose of honorification or beautification. Semantic understanding of the preceding or surrounding text potentially becomes necessary to determine the function of the prefix.

(12) Short *o-noun* sentence
o-uchi desuka.
"The/A/Your house/home?"

4.2.5 The *-rare sonkeigo* form

The potential for ambiguity has been noted as a problem for the honorifics processes above, and this is also the case for the *-rare* suffix, shown in (5), page 20. Shibatani (1991, p. 375) describes the suffix as being homophonous with the suffix for the passive, potential, and spontaneous forms. Word level analysis is clearly insufficient for correct identification, and the potential for ellipsis can again make even sentence level analysis insufficient, as shown in (13). From the context it might be obvious that the teacher did the scolding, but it is also possible to interpret the sentence as the teacher being the one

that was scolded.

(13) Subject honorific form with ellipsis
Sensei wo shikarareta.
"The teacher scolded (somebody)." (honorific)

4.2.6 Suppletive forms (*sonkeigo* and *kenjougo*)

The honorific suppletive forms are used instead of less respectful forms. Being different and limited in number, the usage of these forms should be fairly simple to identify. Because there are different forms for *sonkeigo* and *kenjougo*, it should even be possible to identity the type of honorifics that is being used.

4.2.7 Pronominals and sentence ending particles

The first person pronominals in Japanese provide information about formality, and as single words are fairly easy to identify. Furthermore, pronominal choice can potentially provide information about the sex of the speaker or writer.

Sentence ending particles are similarly easy to identify, being located at the end of sentences, and they can be interpreted as signs of informal or colloquial language.

4.3 Discussion

Above, we examine the feasibility of analyzing Japanese honorifics. We initially consider the possibility of classifying a sentence based on the presence or absence of two independent types of honorifics, namely those that Shibatani (1991) describes as the speaker-addressee axis and the speaker-referent axis. Ideally, it would be possible to use the presence or absence of honorifics

processes to classify arbitrary sentences as belonging to one of the four quadrants created by these two axes, and to use this to infer the genre of the text. Unfortunately, practical problems make this difficult.

Which of the *teineigo* polite and plain forms is used in a sentence can usually be easily determined, but the meaningful absence of *sonkeigo* and *kenjougo* honorific forms is more difficult to establish. The problem is further complicated by the difficulty of reliably identifying even the presence of many of these forms. The result is that even if a text can be placed in either the upper or lower part of Figure 4.1 based on *teineigo* forms usage, it is not possible to say with certainty if it belongs on the left or right side.

In the discussion above, we have primarily considered the possibility of ambiguity or lack of knowledge, not the likelihood. In practice, it might be possible to correctly identify many of the honorifics processes sufficiently often for the possibility of ambiguity to not be a problem. For example, Maeda et al. (1988) describe a way of parsing the *-rare* and *o-verb suru* honorific processes. In many cases, it might also be possible to determine missing information. A way of analyzing a conversation in order to identify the topic, object, or subject in sentences where these elements are not explicitly stated is examined by Yoshimoto (1988). The author is able to frequently identify these elements, but there are cases when this information cannot easily be deduced (see Yoshimoto, 1988, p. 1), a problem also noted in Shirado et al. (2006, p. 405).

For these reasons we focus on the honorific processes that can be identified in a more predictable manner, and primarily analyze the last part of sentences. This is where it generally can be determined if plain or polite *teineigo* forms are used, and where any sentence-final particles will be located. There are additional elements that can be identified if they exist, such as first person pronominals and suppletive forms, and this information can be interesting, but we do not attempt to use it as a basis for the type of genre classification for which Figure 4.1 could be used.

4.4 Summary

This chapter studies the various honorifics processes described in Chapter 3, in order to determine the difficulty of identifying honorifics usage in arbitrary Japanese sentences with a computer.

Ambiguity makes many of the subject and object honorific processes difficult to identify in a reliable manner and for this reason we focus primarily on addressee controlled honorifics and analysis of the last part of sentences in our study of Wikipedia. The classification system we use for this purpose is described in Chapter 5.

Chapter 5

Classification

This chapter proposes a way of classifying honorific processes, and other genre related language characteristics, based on the overview of honorifics processes in Chapter 3, and the analysis feasibility discussion in Chapter 4.

5.1 Purpose

The purpose of the classification system is to aid in the identification and description of the genre of Japanese texts. As discussed in previous chapters, there are are a wide range of genre characteristics, and a classification system will need a way to identify the absence or presence of these characteristics. By applying the system to a Japanese text it should be possible to compare the classified genre characteristics of the text to those of other known genres. In our case, the classification systems needs to be usable for describing characteristics of language found in Wikipedia.

5.2 Related work

Other researchers have used a similar approach to analyze Japanese. Mayumi (2002) analyses transcribed conversations between unacquainted people, in

order to examine the effects of age and gender on politeness. Analysis is done on the sentence level, with three types of characteristics examined: speech level, sentence-final speech level, and utterance type (see Mayumi, 2002, p. 56). The first set of characteristics relates to the sentence level and divides each sentence into four categories depending on whether it contains *super-polite* forms (subject and object honorifics), polite forms (*-masu/desu*), plain forms, or no politeness markers. These categories are named *S*, *P*, *N*, and *NM*, respectively. The sentence-final classification has three categories; *P*, *N*, and, *NM*, depending on whether a sentence has polite forms (*-masu/desu*), non-polite forms, or no politeness markers. The utterance type has four categories; incomplete utterance (I), reversed utterance (R), word level utterance (W), and complete utterance (C). The first covers sentences that are incomplete grammatically, the second sentences where the predicate does not come at the end of the sentence, the third sentences that only contain one word or are ended by a substantive, and the fourth covers all other sentences. In addition, discourse level categorization is used, with marking of topic initiation and shifts in speech levels. So-called back-channel utterances used to indicate understanding or listening are also marked.

Parts of this system is usable with the classification approach we arrive at in Section 4, which focuses on the sentence-final part, but also considers the presence of honorific forms that can appear in other parts of a sentence. This is however a system that is used with manual sentence analysis. As we discussed in Chapter 4, it would be difficult to identify potentially ambiguous honorific processes with a computer. Some additional characteristics can be categorized with the system: the sentence type, topic initiation, and speech levels. The first of these would likely be useful and would require sentence level analysis. The second, for marking of places with topic initiation, would likely require semantic understanding of a sentence, but would also seem to be more useful for analyzing conversations than for the genre characteristics of Wikipedia. The third, for marking of places with changes between e.g.,

plain and polite forms, would be possible to perform, but again would seem to be of more use in conversation analysis. Overall, the system is somewhat limited, having only a small number of categories.

A similar system is used by Musteric (2003, p. 165), in order to determine usage *desu/-masu* polite forms and similar honorific processes. In this system, sentences are considered to consist of two parts: *go*, concerned with the informational content in the sentence, and *watai*, being the part which is used to express attitude. Each of these parts can have one of three speech levels, relative to neutral expressions without any particular degree of formality or informality. These neutral expressions are marked as *0 level*. More polite expressions as marked as *+1* and less polite, or rude, expressions as *-1*. In this system, Musteric (2003) treats the *sonkeigo* and *kenjougo* forms as *go*, *+1*. The *teineigo -masu/desu* forms and the use of the *o-* prefix in front of nouns are treated as *watai*, *+1*. For full text level analysis, ratios are calculated based on the total number of sentences, and the number of sentences with a given *go* or *watai* speech level. A text is given an additional score based on a point system for honorific forms, with points given based on the honorific elements occuring in the text, and their perceived degree of politeness.

This second system uses a slightly different structure, but has some similarities with the one used by Mayumi (2002). The presence of *teineigo*, *sonkeigo*, and *kenjougo* honorific processes affects how a sentence is classified, but the use of numeric values gives a greater scope for differentiating between expressions with different degrees of honorifics usage. The system is also able to identify the so-called *minus* keigo expressions, through the use of negative values. A potential problem is the classification of sentences that contain both honorific forms and colloquial language; in equal number they would cause a sentence to be classified as neutral. The use of a separation between content and attitude makes sense when considering that honorifics processes provide alternative ways of communicating the same information,

Element	Alternatives
1st pers. pron.	*ore*, *boku*, *watakushi*
Verb form	*au*, *aimasu*, *o-ai-shimasu*, *o-me ni kakarimasu*
3rd pers. pron.	*aitsu*, *kare*, *ano kata*, *ano o-kata*
Particle	yo, (none)

Table 5.1: Example expression variations

but the definition of what belongs in each of the two categories seems somewhat awkward.

5.3 Classification system evaluation

To examine the extent to which these two classification systems provide an useful differentiation between expressions of different politeness and formality, we have applied both systems to a set of sentences spanning from the vulgar to the very polite. The sentences are taken from Shibatani (1991, p. 377), and are listed in Figure 5.1. The sentences represent variations on spoken Japanese, and are not necessarily representative for the language variations found in written language, but the same underlying challenge exists for any attempt to characterize a sentence; finding a way to represent the unique characteristics that result from choices having been made among a set of different expression alternatives.

Each sentence essentially contains the same information, what differs is the politeness and formality with which it is said. The primary elements found in the sentences is a first person pronoun, a way of referring to a third person, the verb *to meet*, and a sentence-final particle (in some of the sentences). A summary of the different variations that occur in the sentences is given in Table 5.1.

Each sentence contains a politeness marker, so for the first system there are three categories available for classification of a sentence, and two categories for classification of the sentence-final part. As for the other categories,

(14) a. Vulgar
ore aitsu ni au yo.
"I'll see that fellow."

b. Plain, informal
boku kare ni au yo.
"I'll see him."

c. Polite, informal
boku kare ni aimasu yo.
"I'll see him."

d. Polite, formal
watakushi kare ni aimasu.
"I'll see him."

e. Polite, formal, object honorific
watakushi kare ni o-ai-shimasu.
"I'll see him."

f. Polite, formal, object honorific, honorified 'he'
watakushi ano kata ni o-ai-shimasu.
"I'll see that person (yonder)."

g. Polite, formal, super object honorific, super-honorified 'he'
watakushi ano o-kata ni o-me ni kakarimasu.
"I'll be humbly involved in the eye's (seeing) that honorable yonder."

Figure 5.1: Politeness level sentence examples from Shibatani (1991)

Ex.	Sentence speech level	Sentence-final speech level
a	N	N
b	N	N
c	N	P
d	P	P
e	S	P
f	S	P
g	S	P

Table 5.2: Mayumi (2002) classification system results

all sentences fall into the *complete utterance* category; the other categories are of less relevance for a single sentence. The examples contain elements matching all categories, as shown in Table 5.2. To have as much variation between the example sentences as possible, we have chosen to define *watakushi* as P (polite), rather than S (super polite). The result is still that many sentences get the same classification. Of seven sentences, there are only four different categorization combinations. This kind of classification system gives a general indication of the speech level but is too limited to provide much detail.

The second system provides greater room for differentiating between different degrees of politeness and respect, and we utilize this by giving different politeness scores to the alternatives listed in Table 5.1. The description of the classification system is not complete, and it is unfortunately not obvious how all elements in the sentences fit into it, so we make the following definitions. We consider the presence of sentence-final particles to belong to the *watai* category, along with the choice of first person pronoun. The choice of third person pronouns we define as belonging to the *go* category. The description of the system does not cover all the situations that occur in the example, but we summarize the score values and use the sum to indicate whether the sentence is a *+*, -, or *0* sentence.

The results are shown in Table 5.3, along with the sums. If only the re-

	watai			*go*			
Ex.	1. pp.	*teineigo*	*sum*	Honorifics	3rd pp.	Part.	*sum*
a	-2	0	-2 (-)	0	-1	-1	-2 (-)
b	-1	0	-1 (-)	0	0	-1	-1 (-)
c	-1	+1	0 (0)	0	0	-1	-1 (-)
d	+1	+1	+2 (+)	0	0	0	0 (0)
e	+1	+1	+2 (+)	+1	0	0	+1 (+)
f	+1	+1	+2 (+)	+1	+1	0	+2 (+)
g	+1	+1	+2 (+)	+2	+2	0	+4 (+)

Table 5.3: Musteric (2003) classification system results

sulting categories are considered, there are several sentences with an identical classification, but if only the sums are used, the result is unique when both *go* and *watai* categories are considered. The system is however insufficiently defined for it to be used as is.

5.4 Discussion

A problem with both of these systems is that the categories they define are limited. The reason for this is understandable, because apart from having only analysis of formality level in speech as a goal, they have been designed for manual analysis and classification. Having a small number of categories and an easy way of categorizing expressions reduces the likelihood of human error, but this is not a problem when a computer is used. As noted in Section 5.1, part of the purpose of the classification system is to be able to describe the genre of Wikipedia, and for this purpose the limited number of different classification categories in these systems is insufficient. The way the second system is actually used by the author (see Musteric, 2003, p. 169) does however provide a good starting point for a suitable classification system. The sentences in the text are analyzed, and use of forms such as *-masu/desu* are identified and counted. The number of these forms relative to the total

number of sentences is then calculated.

In Chapter 3, the characteristics of several different genres are described. Many of these involve a choice between alternatives, such as the choice between plain and polite *teineigo* forms, or for the copula, ellipsis or use of the *da*, *desu*, or *de aru* forms. As discussed in Chapter 4, there might be some forms that cannot reliably be automatically detected in a text, but even if these forms are ignored, there will still be several genre characteristics that can be classified.

The first step in a classification approach can therefore be to simply identify the genre characteristics that can be identified, or that the identification system is able to identify. This should give an unstructured overview of the forms that are most common. For the second step, the information on category alternatives can be used to group the information. For example, in the case of the copula, the frequency of the different copula forms can be compared to determine which forms are preferred over other forms. A third and final step would be to compare the set of preferred forms to that of known genres. In the case of newspapers, this might include the use of plain forms and the *de aru* copula form, and in the case of an analyzed newspaper article, the use of plain forms and *de aru* copula forms would be expected to be higher than the alternatives. This approach will allow the results to be analyzed on several levels, and in the case of unexpected results, provides sufficiently detailed information to serve as a good starting point for further analysis.

To illustrate how this classification system would be used, we apply it to a text, consisting only of (14a) and (14b). With only two sentences, the resulting text is not very long, but the same approach can be used on a longer text. Table 5.4 shows the result of analyzing the relevant elements in the two sentences. A descriptive set of categories are used in the table, covering the primary characteristics that occur. The most frequently occurring characteristic is the use of plain verb forms at the end of the sentences,

Characteristic	Number
Plain sentence-final forms	2
Colloquial speech sentence-final particle	2
First person pronoun *ore*	1
First person pronoun *boku*	1
Colloquial third person pronoun	1

Table 5.4: Example expression variations

and a sentence-final particle primarily found in colloquial speech. This kind of enumeration represents the first part of the classification approach. The second step is to group the categories based on the possible alternatives. In the case of sentence-final forms, use of polite forms is an alternative, but they are not used in any of the sentences, making use of plain forms consistent. For first person pronouns, there are several alternatives, and two of these are used once. For this example, the third step would be fairly straightforward because the two examples make consistent use of forms that are found in spoken, colloquial language; a definition of this genre should have a good match with the classification results found in the example.

5.5 Summary

This chapter examines two systems for classifying the level of formality and politeness in Japanese sentences. The systems are applied to a set of example sentences and the results are used as the basis for defining a classification system for Wikipedia. The results of using this classification system are presented in Chapter 8.

Chapter 6

Automated sentence analysis

In Chapter 4, we discussed the identification of characteristics of Japanese genres, based on the analysis of sentences and sentence elements. In Japanese, sentence and word analysis is made complicated by the lack of separation between words. This chapter examines issues related to the type of morphological analysis that will generally need to be performed before any higher level analysis.

6.1 Terminology

Japanese is a non-segmented language and sentences will typically consist of a sequence of characters without any separation between them. Morphological analysis is used to identify the morphemes in these character sequences, and to combine one or more characters into larger units or *tokens*. This operation is called *tokenization*, or *segmentation*. In the case of word sized tokens, the term *word segmentation* is sometimes used (see Utsuro et al., 2000, p. 111), but tokens can consist of smaller units, such as morphemes (see Dridan and Baldwin, 2007, p. 334).

A related, higher level operation, is *bunsetsu identification*, also called *bunsetsu segmentation* and *chunking*. The term *bunsetsu* refers to units that,

for English can be said to be closer to phrasal units than to words (see Murata et al., 2000, p. 1). One example of a single bunsetsu can be a noun and particle pair. Chunking identifies bunsetsu in a sentence and is performed after part-of-speech tagging (see Utsuro et al., 2000, p. 111). The identified bunsetsu can then be used for *dependency analysis*, which is concerned with the determination of the relationship between different bunsetsu.

6.2 Tokenization

The correctness of any analysis of elements in a sentence depends on the correctness of the morphological analysis, making an understanding of the accuracy and potential problems of this operation necessary for any higher level analysis. There are several honorifics processes and related genre characteristics, as described in Chapter 3, but despite having limited the information that we are trying to identify to that found in the last part of a sentence, it is still necessary to first perform a morphological analysis in order to separate the various sentence elements.

6.2.1 Analysis example

Morphological analysis can be done using a dictionary with information about word classes, and a description of the syntax of Japanese, including information on inflection patterns. An example sentence is shown in (15), taken from Makino and Tsutsui (2002a, p. 37). The sentence is shown both as it would be written in Japanese and how it can be tokenized[1].

[1] As there is no single well defined procedure for tokenization, this is just an example.

(15) Japanese sentence (honorific)
先生はアメリカの大学で日本語を教えられます。
先生 は アメリカ の 大学 で 日本語 を 教え られ ます 。
sensei ha amerika no daigaku de nihongo wo oshieraremasu.
"The professor will teach Japanese at an American college."

In this example, the sentence starts with the noun *sensei*, which can be found by looking up the first two characters in a dictionary. A noun will typically be followed by a particle, as is the case in this example, allowing the third character to be classified. Both of the two first characters might have separate entries in a dictionary, but an analysis might determine that two separate nouns with no particle in between them is less likely to be correct than one longer noun with a particle following it[2]. A similar procedure can be applied until the final word in the sentence, where the verb inflection patterns need to be taken into consideration because a normal dictionary will not contain an entry for the full verb form in the example. The dictionary form is *oshieru* (to teach), which does not occur in the sentence but can be derived from the inflected form. In this example, the honorific *-rare* suffix and the polite *-masu* form have been separated from the verb stem.

A byproduct of this procedure is that the word class of each element is known, because this information is required to separate the sentence elements. The result is that word level analysis becomes simple at this point. Vocabulary usage, sentence-final particles, pronoun types, and suppletive forms can be identified. This is also a good starting point for doing more advanced analysis of a sentence, such as trying to determine if, as in this case, the *rare* suffix is honorific or not.

[2]These decisions can done based on probabilities generated from Japanese corpora (see Asahara and Matsumoto, 2004). In the case of ambiguities, a decision might not be taken before the entire sentence has been analyzed.

6.2.2 Problem areas

At first glance, the morphological analysis would seem to be fairly straightforward to perform even with a computer. Combine an encoding of inflection patterns with a dictionary and the result should be a tool that is able to perform the morphological analysis required to separate the elements in a sentence. The small number of irregular verbs in Japanese would make the information required to do this relatively modest, but in practice, there are problems that make morphological analysis nontrivial.

An obvious problem with a segmentation approach that relies on information obtained from a dictionary or corpus is the handling of unknown words (see Asahara and Matsumoto, 2004). A system should ideally be able to both correctly perform word segmentation and do part-of-speech tagging for unknown words, but this becomes difficult if the segmentation procedure depends on words being known.

Lexeme identification faces a challenge in that even the same form of a single lexeme can be written in different ways. An example given by Den et al. (2008) is the verb *arawasu* (to express), which can be written using only hiragana ("あらわす"), or in two different ways using kanji ("表す" and "表わす"). With four different writing systems (kanji, hiragana, katakana, and romanji) and the possibility of having more than one writing system used in a single word, there is ample room for having multiple ways of writing any given lexeme (see Kacmarcik et al., 2000). A related problem is that of multiple lexemes being written the same way.

In some cases, there might additionally be ambiguity with regards to the segmentation of words (see Kudo et al., 2004). For example, "東京都" can be interpreted as *higashi kyouto* (east Kyoto), or *tokyou bu* (Metropolis of Tokyo). Having the possibility of writing words both with and without kanji can further complicate this problem by increasing the room for ambiguity. Kudo et al. (2004) mention kanji such as "内" that can have the reading *nai*, which is identical to one form of an auxiliary verb if the kanji is written in

hiragana.

The operation of a tokenizer, as described in Section 6.2.1, is fairly simple, but practical issues such as those described above lead to the creation of a tokenizer not being a simple process if the goal is to support a wide range of possible Japanese sentences. Fuchi and Takagi (1998) describe the need for complicated fine-tuning and statistical analysis for many existing systems. We initially made an attempt to create a simple tokenizer suited for studying Wikipedia, but quickly abandoned that approach in favor of using an existing system. A significant amount of research has been done in this problem area and tokenization appears to have become a fairly mature field. Existing systems can achieve an accuracy of over 97% (see Den et al., 2008).

6.2.3 Tokenizer comparison

Because we abandoned our ambition of writing a similar program, we needed to choose a tokenizer that could be used to segment the text in the Wikipedia articles. The publicly available morphological analyzers that appear to be most commonly used are *ChaSen*[3], *MeCab*[4], and *JUMAN*[5].

We did a simple performance comparison of the three tokenizers, using a 33.1 MB *EUC-JP* encoded file with text from Wikipedia. Each tokenizer was fed the input file, while the total processing time was measured. The tokenizer output was discarded to avoid any influence on the measured time from write operations to the hard disk. The tests were made on an *Ubuntu 8.10 Linux* system with an 1.83 GHz *Intel Core Duo CPU*. Each test was repeated three times, and the median value for the total processing time was used. There were only minor variations in processing times. A summary of the results is shown in Table 6.1, along with the tokenizer version numbers.

Of the three tokenizers, MeCab is the fastest, using only 27.48 seconds.

[3] `http://chasen-legacy.sourceforge.jp/`
[4] `http://mecab.sourceforge.net/`
[5] `http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/juman-e.html`

System	Version	Median processing time (sec)
MeCab	0.97 (IPADIC)	27.48
ChaSen	2.4.3 (IPADIC)	40.94
JUMAN	5.1	490.65

Table 6.1: Tokenizer processing times

ChaSen is slightly slower, at 40.94 seconds, while JUMAN is almost ten times as slow as ChaSen. In an experiment done by Fuchi and Takagi (1998), the processing time of JUMAN was found to be almost twice that of ChaSen, which indicates that either the performance of ChaSen has improved, or there is something in our configuration which has a negative impact on the performance of JUMAN[6]. Whichever of these possibilities is correct, JUMAN is clearly the slowest of the three. These tests were performed with the text file encoded in EUC-JP, while Wikipedia is encoded in *UTF-8*. Running MeCab on an UTF-8 encoded version of the same file resulted in only a slight increase in processing time, at 30.18 seconds.

We did not perform any additional tests, but according to Den et al. (2008), the accuracy of MeCab is higher than that of ChaSen, at above 98%, which would seem to make the choice of tokenizer easy. JUMAN segments sentences into smaller morphemes than ChaSen (see Sasano and Kurohashi, 2008), but the difference in performance between JUMAN and MeCab makes the cost of these potential benefits quite expensive. Performance is an important factor in our scenario because Wikipedia contains a large amount of text. We used MeCab for our analysis work.

[6]No special options were used with any of the programs, with the exception of MeCab, which was started with the *-b 32768* option to avoid warnings about long lines. Running without this option did not appear to have any impact on performance.

Tokens	Part-of-speech tagging
先生	名詞,一般,*,*,*,*,先生,センセイ,センセイ
が	助詞,格助詞,一般,*,*,*,が,ガ,ガ
太郎	名詞,固有名詞,地域,一般,*,*,太郎,タロウ,タロー
を	助詞,格助詞,一般,*,*,*,を,ヲ,ヲ
叱ら	動詞,自立,*,*,五段・ラ行,未然形,叱る,シカラ,シカラ
れ	動詞,接尾,*,*,一段,連用形,れる,レ,レ
た	助動詞,*,*,*,特殊・タ,基本形,た,タ,タ
。	記号,句点,*,*,*,*,。,。,。
EOS	

Table 6.2: MeCab output for (5)

6.2.4 MeCab based tokenization

In Chapter 4 and Chapter 5, we discussed problems related to the identification and classification of honorifics processes in Japanese. These operations can only be performed after tokenization, and require tokenization to have been done in a way which provides sufficient information for classification. To examine the extent to which this is possible with MeCab, we used it to tokenize some of the example sentences we have studied in previous chapters.

Table 6.2 shows the MeCab output for (5), on page 20. The tokens are identified as nouns (“名詞”), particles (“助詞”), verbs (“動詞”), and auxiliary verbs (“助動詞”). In the case of the verbs, information is provided on inflection; the token (“叱ら”) is correctly identified as *mizenkei* (“未然形”) of a *godan* (“五段”), *ragyou* (“ラ行”) verb. The rightmost part identifies the lexemes, “叱る” in the case of the verb. The term “EOS” is used to signal the end of the sentence.

This sentence uses the *-rare* suffix, indicating respect towards the subject of the sentence (“先生”). MeCab marks “れ” as a verb and a suffix (“接尾”). It is additionally classified as being of type *ichidan* (“一段”), and in the *renyoukei* (“連用形”). Note that it is not marked as having a honorific, passive, or similar function; identification of these functions is generally not

Tokens	Part-of-speech tagging
先生	名詞,一般,*,*,*,*,先生,センセイ,センセイ
は	助詞,係助詞,*,*,*,*,は,ハ,ワ
刺身	名詞,一般,*,*,*,*,刺身,サシミ,サシミ
を	助詞,格助詞,一般,*,*,*,を,ヲ,ヲ
食べ	動詞,自立,*,*,一段,未然形,食べる,タベ,タベ
られ	動詞,接尾,*,*,一段,連用形,られる,ラレ,ラレ
た	助動詞,*,*,*,特殊・タ,基本形,た,タ,タ
。	記号,句点,*,*,*,*,。,。,。

Table 6.3: MeCab output for (16)

attempted by a morphological analyzer. The output from a tool like MeCab can serve as the starting point for this kind of analysis, but the way in which the honorific form is constructed for different verb types makes this task slightly more difficult.

(16) Plain form
先生は刺身を食べられた。
sensei ha sashimi wo taberareta.
a. "My teacher ate sashimi." (honorific)
b. "My teacher could eat sashimi." (potential)
c. "Someone ate sashimi and my teacher was unhappy." (indirect passive)

In Table 6.3, the MeCab output of (16) is shown, a similar sentence with the *-rare* suffix, taken from Makino and Tsutsui (2002a, p. 369). In this case, the verb ("食べる") is of type *ichidan* ("一段") and the verb stem does not end with *a*. With this type of verb, the full *-rare* suffix comes as a separate morpheme after the verb stem. Again, the suffix is not marked as honorific or passive. However, to recognize the sentence as potentially being honorific, it is possible to look for a verb in *mizenkei* form, followed by a *-re* or *-rare* suffix, depending on verb type. Determining whether, for example, (16) is honorific, potential, or indirect passive is, as discussed in Chapter 4,

Tokens	Part-of-speech tagging
食べ	動詞,自立,*,*,一段,未然形,食べる,タベ,タベ
られ	動詞,接尾,*,*,一段,連用形,られる,ラレ,ラレ
まし	助動詞,*,*,*,特殊・マス,連用形,ます,マシ,マシ
た	助動詞,*,*,*,特殊・タ,基本形,た,タ,タ
。	記号,句点,*,*,*,*,。,。,。

Table 6.4: MeCab output for polite version of (16), past

Tokens	Part-of-speech tagging
食べ	動詞,自立,*,*,一段,未然形,食べる,タベ,タベ
られ	動詞,接尾,*,*,一段,連用形,られる,ラレ,ラレ
ます	助動詞,*,*,*,特殊・マス,基本形,ます,マス,マス
。	記号,句点,*,*,*,*,。,。,。

Table 6.5: MeCab output for polite version of (16), non-past

more problematic, but these problems need to be addressed by higher level analysis, and are not caused by MeCab.

Analysis of the other honorific processes can be done in a similar way. The MeCab output from the analysis of the polite past *-masu* form of the verb in (16) is given in Table 6.4. The output is identical to that in the last part of Table 6.3, with the exception of one extra auxiliary verb; *mashi* (“まし”), the past form of *masu*. The non-past form of the verb is shown in Table 6.5, and the auxiliary verb is again segmented as a separate token, here *masu* (“ます”). The forms are different, but the lexeme is identified in both examples as *masu* (“ます”). For the two honorific passive sentences this is not the case. In *shikareta*, the token is identified as *reru* (“れる”), while for *taberareta*, it is *rareru* (“られる”). MeCab tokenizes the sentences, but for tasks such as honorific process identification, it is still necessary to do additional processing that requires knowledge of Japanese syntax.

An interesting case is that of (14g), on page 41, the tokenization of which is listed in Table 6.6. The sentence contains two words with the *o-* prefix; *o-kata* (“おかた”) and *o-me* (“おめ”), but only the first is tokenized with

Tokens	Part-of-speech tagging
わたくし	名詞,代名詞,一般,*,*,*,わたくし,ワタクシ,ワタクシ
あの	フィラー,*,*,*,*,*,あの,アノ,アノ
お	接頭詞,名詞接続,*,*,*,*,お,オ,オ
かた	名詞,接尾,一般,*,*,*,かた,カタ,カタ
に	助詞,格助詞,一般,*,*,*,に,ニ,ニ
おめにかかり	動詞,自立,*,*,五段・ラ行,連用形,おめにかかる, オメニカカリ,オメニカカリ
ます	助動詞,*,*,*,特殊・マス,基本形,ます,マス,マス
。	記号,句点,*,*,*,*,。,。,。

Table 6.6: MeCab output for polite version of (16), non-past

the *o-* prefix as a separate element. The reason appears to be connected to the dictionary used by MeCab, because with the JUMAN dictionary *o-kata* appears as a single word. Most likely, this is caused by the whole expression *o-me ni kakaru* (“おめにかかる”) existing as a single entry in the dictionary, while *o-kata* only exists as a word in the JUMAN dictionary, where the lexeme is identified as “御方”. Other expressions, such as *o-sake*, are tokenized with *o-* and *sake* as separate tokens by all three tokenizers. Reliably detecting the o-prefix would require either the *o-* entries in the dictionary to be removed, or the creation of a list of exceptions containing the *o-* prefixed lexemes that do appear in the dictionary. Because this would involve an examination of the entire dictionary, we elected to only note that any analysis of the *o-* prefix would possibly be incomplete.

Another dictionary related issue we discussed in Section 4.2.3, is that of nouns that are identical to the adverbial form of verbs, such as *tasuke* in the honorific form *o-tasuke shita*. The MeCab output for this expression is shown in Table 6.7, with *tasuke* (“助け”) classified as a noun (“名詞”). The final part of the output for a similar sentence, taken from Makino and Tsutsui (2002a, p. 40), is shown in Table 6.8. The verb in this sentence is *to carry*, or *motsu* (“持つ”), and is in the same form as *tasuke*, but here it is classified as a verb (“動詞”). Both adverbial forms are classified as verbs if the JUMAN

Tokens	Part-of-speech tagging
お	接頭詞,名詞接続,*,*,*,*,お,オ,オ
助け	名詞,一般,*,*,*,*,助け,タスケ,タスケ
し	動詞,自立,*,*,サ変・スル,連用形,する,シ,シ
た	助動詞,*,*,*,特殊・タ,基本形,た,タ,タ
。	記号,句点,*,*,*,*,。,。,。

Table 6.7: MeCab output for honorific form *o-tasuke shita*

Tokens	Part-of-speech tagging
お	接頭詞,名詞接続,*,*,*,*,お,オ,オ
持つ	動詞,自立,*,*,五段・タ行,基本形,持つ,モツ,モツ
し	動詞,自立,*,*,サ変・スル,連用形,する,シ,シ
ます	助動詞,*,*,*,特殊・マス,基本形,ます,マス,マス
。	記号,句点,*,*,*,*,。,。,。

Table 6.8: MeCab output for honorific form *o-mochi shimasu*

dictionary is used, but a detailed examination of the dictionary would again be needed to guarantee that this is consistent for other verbs.

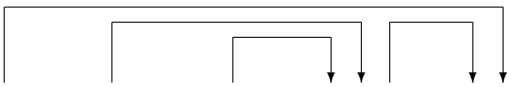
6.3 Chunking and dependency analysis

The output from a tokenizer like MeCab can be used for bunsetsu chunking, which is an operation generally performed before any higher level analysis (see Murata et al., 2000). Operations such as dependency analysis typically operate on bunsetsu rather than the words or morphemes in a sentence. An example of a sentence with identified bunsetsu, taken from Murata et al. (2000), is shown in (17). The *"/"* character shows the division between the different bunsetsu in the sentence.

(17) Identified bunsetsu
boku ga / bunsetsu wo / matomeageru.
"I identify bunsetsu."

The bunsetsu chunking operation can be performed with a fairly high degree of accuracy. Murata et al. (2000) report correctness of over 98% with one possible technique. The bunsetsu are not necessarily directly useful by themselves, but these numbers indicate that it should be possible to perform bunsetsu chunking without adversely affecting the accuracy of tasks that operate on bunsetsu.

An example of dependency analysis is given in (18), taken from Kudo and Matsumoto (2002). The relationships between the bunsetsu in the sentence is shown by the arrows. For example, *warm* modifies *heart* and *he* modifies *be moved*.

(18) 

彼は｜彼女の｜温かい｜真心に｜感動した。

he her warm heart be moved

kare ha kanojyou no atatakai magokoro ni kandoushita.

"He was moved by her warm heart."

In Chapter 4, we discussed several possible situations where ellipsis could cause ambiguity on the sentence level. This kind of dependency analysis would be a useful starting point for doing full text level analysis, in order to attempt to identify the omitted elements in ambiguous sentences. Dependency analysis does however appear to be somewhat more difficult than tokenization and bunsetsu chunking. The success ratios reported in published research lie at around 90% for the number of correct dependencies, and 50% for the number of sentences with all dependencies correctly identified (see Murata et al., 2000; Tamura et al., 2007; Imamura et al., 2007).

We argue in Section 4.3 that ambiguity would make the identification of many of the honorific processes unreliable. The possibility of incorrect dependency analysis would further complicate this process. Because of the difficulty and unreliability of this kind of analysis we chose to concentrate on

simpler and more reliable word and sentence level analysis of the Wikipedia content.

6.4 Summary

This chapter covers practical issues related to computer based analysis of Japanese sentences, using tools such as the morphological analyzer MeCab. Division of sentences into morphemes with this kind of tool can be done with a high degree of accuracy, but additional processing is required to identify honorific forms. In some cases, the output from MeCab is not consistent, complicating post-processing. Higher level-analysis, such as dependency analysis, quickly becomes difficult to do accurately, but we do not make use of this kind of analysis in our examination of Wikipedia.

Additional challenges related to processing the content of Wikipedia are discussed in Chapter 7.

Chapter 7

Preliminary analysis

In this Chapter, we describe the results of a preliminary analysis of Wikipedia, focusing on what is described in Wikipedia as correct writing style, and how to obtain the Wikipedia article text.

7.1 Expected writing style

Wikipedia is a web-based encyclopedia, and can be accessed with a web-browser, but it is possible to download a snapshot of the data used to build the web-pages. The description below is partly based on the content of Wikipedia articles accessed via a browser, and in these cases we list the title of the article and the last modification date. It is also based on content from a downloaded snapshot marked with the date July 24, 2008[1]. Background information on Wikipedia can be found in Chapter 2.1.

In addition to the main encyclopedic content in Wikipedia, there are separate pages targeted at editors that document various aspects of the project, including the proper writing style for articles. These pages describe how to choose article titles, write dates, etc. For the Japanese Wikipedia, there is

[1]The file *jawiki-20080724-pages-articles.xml*, containing a snapshot of the Japanese Wikipedia, was obtained from: *http://download.wikimedia.org/jawiki/*

especially one page that covers writing style[2]. This article describes the recommended practice for the Japanese version of Wikipedia. The guidelines are, according to the page, approved by many users but are not official policy. The general goal is to achieve consistency and a style which is close to that of printed text. Following the guidelines is recommended, but not an absolute requirement; for some types of articles a different style might be more appropriate, in which case agreement should be reached on the suited style for these articles. In cases where following the guidelines would be an obstruction to good writing, not doing so is acceptable.

A section titled *buntai* (literary style) is most directly relevant to the types of language usage characteristics we are examining in this thesis. The section contains four short points:

- Text should be written using a combination of Chinese and Japanese characters.
- The encyclopedia articles should use direct (plain) forms, including *de aru* and *da*.
- Articles in the "Wikipedia:" and "Help:" name-spaces[3] should use distant (polite) forms, such as *desu* and *-masu*, but this limitation does not apply to lists, etc.
- Lists should make use of short words and sentences.

The first point describes the style of writing common in modern Japanese, and the last point is not relevant for the main article text, making the second and third points the primary documentation of the proper writing style. The description is short and not very detailed, possibly implying that potential

[2]This title of this article in the Japanese Wikipedia is "Wikipedia:表記ガイド". We examined the version of the page last modified May 1, 2009.

[3]Articles prefixed with these keywords contain non-encyclopedic information, such as the document describing these writing guidelines.

editors are assumed to be familiar with these styles. The expected style for the encyclopedia articles is plain forms and use of the *de aru* copula, which corresponds with the writing style generally found in newspapers and scholarly articles (see Section 3.1.1). The short description does however leave a lot of room for variation, for example, with regards to conjunctive verb forms.

Other points of interest in the document concerns the writing style of particles, suffixes, and auxiliary verbs that can be written using both kanji and hiragana. In these cases, the hiragana form is generally recommended, including in the case of the honorific *o-* prefix. Additionally, it is in general not recommended to use hiragana in cases where this might cause a word to be misinterpreted. Use of exclamation and question marks are discouraged, unless these characters are part of proper names or are used in quotations. Finally, colloquial language is regarded as improper.

7.2 Content extraction

For large-scale analysis of Wikipedia, it is more efficient to download a single snapshot of the articles rather than accessing a large number of web pages with a browser. The snapshot files[4] contain an UTF-8 encoded XML file. XML (see Bray et al., 2008) is a markup language suited for computer based processing. As an example, Figure 7.1 shows the XML markup for the article on the Japanese language ("日本語"). The most interesting information can be found in the *title* tag, giving the name of the article, and the *text* tag, where the content can normally be found (the contents of the article have been removed here to illustrate the XML structure). Most of the remaining fields provide information about the revision of the article. Due to the highly structured nature of the document, it is easy to extract the title and text body.

[4]See `http://download.wikimedia.org/` for download information.

```
<page>
  <title>日本語</title>
  <id>11</id>
  <revision>
    <id>20711870</id>
    <timestamp>2008-07-12T05:31:36Z</timestamp>
    <contributor>
      <username>R28Bot</username>
      <id>177951</id>
    </contributor>
    <minor />
    <comment>ロボットによる: 秀逸な項目へのリンク
     [[sco:Japanese leid]]</comment>
    <text xml:space="preserve">

    </text>
  </revision>
</page>
```

Figure 7.1: XML markup for Wikipedia entry on Japanese

An example of an article body is shown in Figure 7.2, again for the entry on the Japanese language. Only a small part of the article is shown, and some lines have been deleted, but the remaining lines illustrate the type of content that can be found in a typical article body; the actual text is only one part of an article. The first part, starting with "`{{言語|name=日本語`", is a so-called *template*, and is used to encode different types of information in a standardized manner. These templates, and other forms of markup, are used when the content is presented in media such as a web browser or book. The actual processing performed depends on the medium, but the markup is used to make it possible to, for example, create links between different articles. This template can be used to describe aspects of a language (here Japanese). The information in the template is presented in a side-box if a browser is used to visit this page on the Wikipedia web site. For our purpose, this information is not interesting, and should be removed. The template is enclosed between a pair of "`{{ }}`" characters, and like this template, can be

nested, but removal is still possible as long as the template is syntactically correct. There is unfortunately no guarantee that this will always be the case, because an editor might, for example, make an error when creating or modifying a template, but any incorrect or incomplete template removal will still be limited to a single article due to each page being encoded separately in the XML file. There are cases where templates might contain full sentences, but because this is not where the main part of the text is contained, we still remove the content of all templates.

The first part of the first paragraph of the text follows the template, but also the text contains markup elements. For example, the "[[]]" characters, such as in "[[日本列島]]", are used to create a link to the page with this title when the page is presented for use on a web server. In some cases, such as for "[[日本|日本国]]", which occurs in the template, the link will be created to the page for "日本" (Japan), while the text to the right of the "|" character is what will actually be shown in a browser, in this case "日本国". We remove this kind of markup and use the text to the right of the "|" character when it occurs. The quote mark characters around the first word have a similar markup function, in this case making the word be shown in a bold typeface when rendered in a browser. Removal of this kind of markup is also unproblematic.

More debatable is the proper way of handling parentheses and quote marks. In this case, the parentheses contain two possible readings of the first word, written in hiragana. Quote marks can basically contain anything, including citations of any type of language, regardless of the writing style of the surrounding text. By removing quoted text we avoid incorrect classification of the text, and the same can be done for text in parentheses. It is possible that this will remove text that correctly represents the writing style of the document, especially in the case of parentheses, but we consider this preferable to incorrect classification. Because this type of punctuation contributes to the structure of a sentence, we leave a marker in the text to

```
{{言語|name=日本語
|nativename={{IPA|n&amp;#690;i&amp;#614;o&amp;#331;&amp;#331;o}}
  ({{IPA|n&amp;#690;ippo&amp;#331;&amp;#331;o}})
|familycolor=#dddddd
|states=[[日本]]など（「[[#分布|分布]]」の節参照）
|region=[[東アジア]]など
|nation=[[日本|日本国]]（事実上）
|agency=特になし&lt;br/&gt;[[文化庁]][[文化審議会]]国語分科会（事実上）
|iso1=ja
|iso2=jpn
|iso3=jpn
|sil=JPN}}
'''日本語'''（にほんご、にっぽんご）は、主として、[[日本列島]]で[[大和民族]]（日本人）によって使用されてきた[[言語]]である。外国から[[帰化]]した者などを除いて、ほぼ全ての日本人は日本語を第一言語とする。

== 関連項目 ==
&lt;div style=&quot;float: left;&quot;&gt;
 [[画像:Wikipedia.png|50px|none|Wikipedia]]&lt;/div&gt;
&lt;div style=&quot;margin-left: 60px;&quot;&gt;
'''日本語版'''の'''[[ウィキペディア]]'''があります。
&lt;/div&gt;&lt;/div&gt;
{{Sisterlinks
|commons=Cattegory:Japanese language
|wikibooks=日本語
|wiktionary=Category:日本語
}}
&lt;!--五十音順に示す。--&gt;
* [[アイヌ語]]
* [[方言]]
** [[日本語の方言]]
* [[協和語]]
* [[現代日本語文法]]
* [[日本における漢字]]
* [[和製漢語]]
* [[和製英語]]
* [[Engrish]]
```

Figure 7.2: Content from Wikipedia entry on Japanese

indicate that the contents of parentheses or quote marks have been removed. As with the templates, there is no guarantee that the sentences in a text body are syntactically correct; the parenthesis characters in a text might not match correctly. For example, in an expression such as "（。。。)", where a Japanese parenthesis is used at the start and a western parenthesis is used at the end, the contents will not be correctly removed. We observed examples of this in the text, but it is not an extensive problem[5].

Below the text paragraph can be seen the expression "== 関連項目 ==", which is a section title. Titles do generally not contain full sentences and we remove them from the text. A set of markup codes which are used when the page is presented in a browser follows the section title. This kind of markup would usually be detectable by the presence of tags similar to those in Figure 7.1, starting with a "<" character and ending with ">" or "/>". However, in this case, they are quoted, with "<", which is a quoted representation of the "<" character. To correctly remove this kind of markup it is necessary to first replace all quoted characters with the actual characters they represent, and then look for tags that can be removed. Interestingly, while the first text paragraph uses plain verb forms (*dearu* and *suru*), the text within the markup uses a polite verb form (*arimasu*). In a browser, the markup is presented as a side-box with a link to information about the Wikipedia project. It would appear that the proper writing style for this type of box is different from that of the main text. The template below the markup is rendered in a way consistent with this observation. It is possible that there are valid sentences contained within this type of markup, but because this is generally not the way the main part of the text is encoded, we remove also this kind of text.

The last part of the text contains a list, with list entries prefixed by "*" characters. Because the guidelines for the Japanese Wikipedia specify

[5]In total, 24,232 sentences (0.37% of all sentences extracted from the file) have a parenthesis character after preprocessing, either due to this type of mismatch or other limitations in the cleaning process.

```
日本語 XpXは、主として、 日本列島 で 大和民族 XpXによって使用
されてきた言語 である。
外国から 帰化 した者などを除いて、ほぼ全ての日本人は日本語を
第一言語とする。
```

Figure 7.3: Processed text output from Wikipedia entry on Japanese

that items in a list should be short words or sentences, essentially making lists have a separate writing style, we remove all list entries in a text. In this example, the list entries are mainly nouns that link to other Wikipedia articles.

After processing, the content from Figure 7.2 is reduced to that in Figure 7.3. All the markup, templates, and lists have been removed, leaving only the main text. The parentheses, and the text inside them, have been replaced with the characters "XpX", which do not occur in the input text, and are used to indicate that parentheses existed at this point in the sentence. After this type of content extraction and cleanup, the text is ready for processing with MeCab.

7.3 Discussion

Many decisions had to be taken with regards to what should be removed and what should be retained from the Wikipedia article data. Our overall goal has been to extract as much text as possible, without having too many exceptions that need special handling during processing. To ensure removal of text with a writing style different from that of the main article content, such as the text in side-boxes, we might have decided to remove more text than necessary. A more comprehensive examination of the Wikipedia markup language might show a way to extract more text from the Wikipedia snapshot files, but we consider the approach we have used to represent a good starting point. Manual inspection was used to verify the correctness of the content

extraction procedure.

7.4 Summary

This chapter describes the result of a preliminary analysis of Wikipedia, with regards to expected writing style and the steps required to prepare the articles for automated analysis. Chapter 8 presents the results of the final analysis, along with a comparison of the observed writing style, and the style recommended by the Wikipedia guidelines.

Chapter 8

Wikipedia classification

In this chapter, we present the results of our analysis of Wikipedia, which is based on preparations done during a preliminary analysis described in Chapter 7. Aspects of the genre of Wikipedia are characterized and described based on the classification procedure in Chapter 3.

8.1 Analysis procedure

The data flow of the automated text analysis is shown in Figure 8.1, with the data flowing from left to right in the figure. A Wikipedia XML snapshot file is taken as input, and preprocessed to remove markup and other non-text elements using the procedure described in Chapter 7. The output is a stream of Japanese sentences, which are given as input to MeCab for morphological analysis. The classification is done by a tool we wrote in *perl* to identify sentence characteristics. Statistical analysis is then performed on the output from this tool, with the results presented below.

The preprocessing step is performed by a second *perl* program we wrote to extract the text from the XML files. It uses a set of rules to detect unwanted content and remove it from the text. It then attempts to identify the sentences in the remaining text and places each sentence on a single line

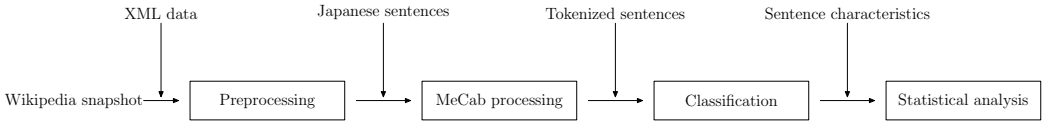


Figure 8.1: Analysis data flow

in the file. Each line is finally classified based on the punctuation on the end of the line. Lines that end with a Japanese full stop character, exclamation mark, or question mark, are treated as valid sentences. The remaining lines are marked to indicate invalidity, and typically contain text without any final punctuation, or unwanted content that our program was unable to identify and remove in the first preprocessing step. In some cases, this is unavoidable; articles containing text like poems or incomplete sentences are difficult to properly analyze, and we remove these lines before processing with MeCab. The extent to which these types of lines occur in the data is examined below.

8.2 Data sets

Our primary goal is to classify the language used in the Wikipedia encyclopedia articles. We obtained this data from the file *jawiki-20080724-pages-articles.xml*, mentioned in Chapter 7. The results from analysis of this file was compared to the contents of the discussion and user pages, where the writing style found in the encyclopedic content is not expected. The text for these pages was obtained from the file *jawiki-20080724-pages-meta-current.xml*, which has the same structure as the main XML file. The analysis steps shown in Figure 8.1 were applied in the same manner to both files.

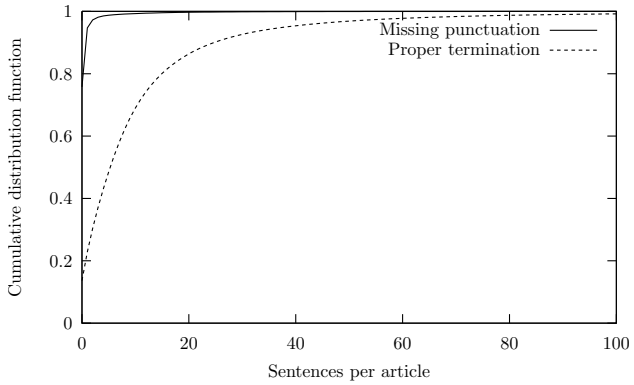
After preprocessing, the encyclopedia data contains $6,510,554$ sentences, the discussion page data contains $697,278$ sentences, and the user page data $1,072,337$ sentences. This is after removal of lines missing a correct punctuation character. It is important to know the efficiency of the content extraction procedure in order to be able to evaluate the classification results. If the ex-

tracted data only represents a small part of the article text, it would be difficult to argue convincingly that the results are representative. Figure 8.2 shows the distribution of the number of properly terminated sentences and invalid lines per article, for the three data sets[1].

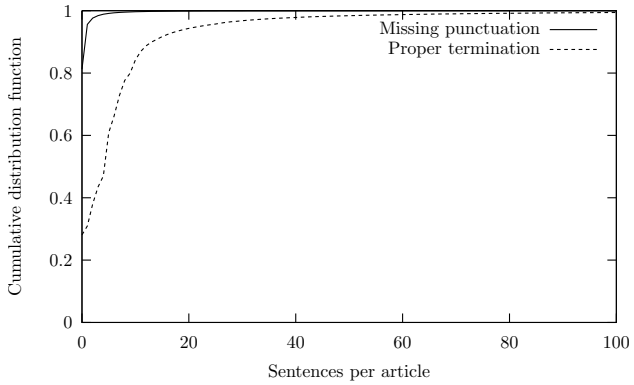
The properly terminated sentences clearly dominate in all three data sets. For the encyclopedia data, only 13.6% of the articles have no sentences with proper punctuation at the end of the line, while as many as 75.8% of the articles have no invalid lines. In total, 94.7% of the articles have no more than one invalid line. The highest number of properly terminated lines in an article is 1748. For the user pages, 28.1% of the articles have no valid lines, and 81.2% have no improper lines. The number of articles with at most one invalid line is 95.6%. The highest number of proper lines in an article is 810. Of the discussion pages, only 2.9% of pages have no valid lines, and 72.6% have no improper lines. At most one improper line is found in 90.3% of the article pages. The highest number of valid lines in a discussion page is 887.

From the numbers above, it can be seen that the number of valid lines dominate. More than 90% of all three page types have at most one invalid line. The percentage of pages lacking any valid lines is not insignificant, being 13.6% for the encyclopedia articles, but there are articles that contain few sentences or primarily have as a purpose to organize information in lists. For the discussion pages, only 2.9% have no valid lines, because discussion pages are likely not created unless an editor wishes to make a comment or carry a discussion with regards to the article the discussion page is created for. The low value is a good indication that the high number of encyclopedia articles with no valid lines is not caused by overly aggressive removal of content during the preprocessing stage. The number of user pages with no content

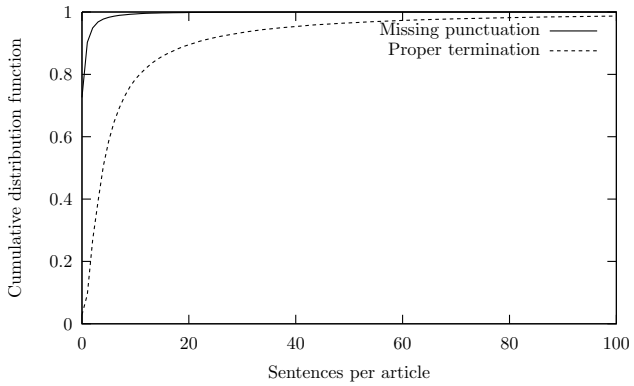
[1]Note that this is after the first unwanted content removal step. A comparison with the contents of the articles before removal of markup might be interesting, but making a meaningful comparison to this data is difficult because it contains non-sentence elements such as lists and markup, which must be removed. See the discussion in Chapter 7 for details.



(a) Encyclopedia articles



(b) User pages



(c) Discussion pages

Figure 8.2: Sentence distribution

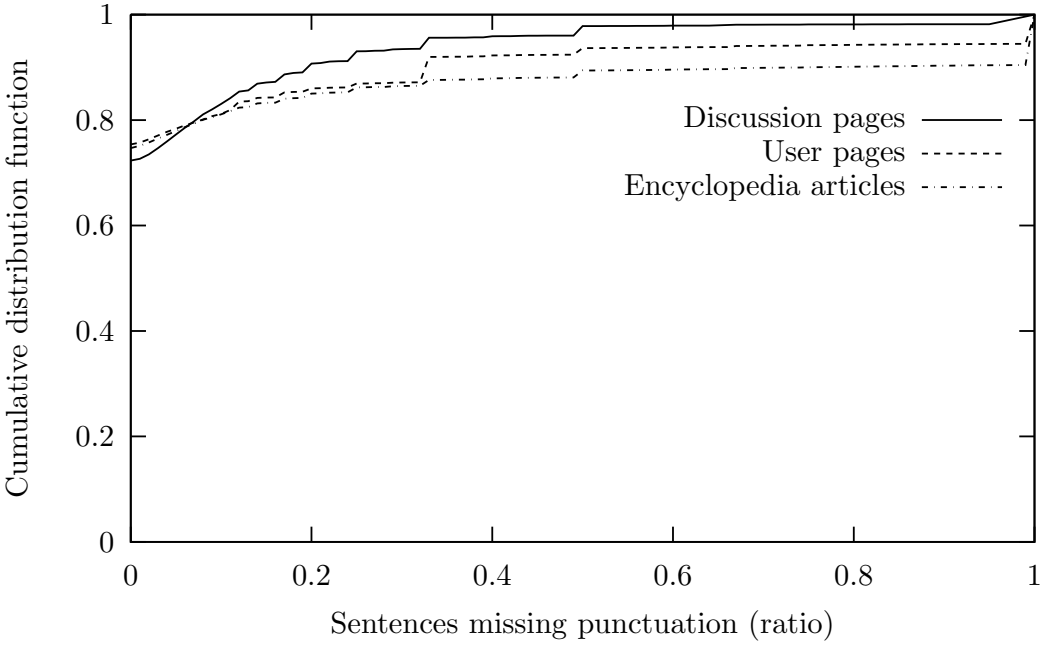


Figure 8.3: Improperly terminated sentence distribution

is slightly higher, at 28.1%, but users are not required to write any text on their own pages, and in some cases the content is only a small number of words without any termination. Some user pages contain text in English or other languages that do not make use of the Japanese writing system.

A comparison of the distribution of ratios of improperly terminated lines in an article, relative to the total number of lines, is given in Figure 8.3. As the figure shows, there is some variation between the three data types but no major differences. The discussion pages have the lowest number of articles with the majority of lines being improperly terminated and this is likely due to these pages primarily containing text. Some of the improperly terminated lines in the discussion pages contain timestamps (typically the date and time a user made a comment) that could have been more thoroughly removed during preprocessing.

Overall, the content extraction is fairly efficient, and the language in the

extracted sentences should give a representative overview of the language used in Wikipedia. There are articles where only a small number of sentences are extracted, but in some cases this is unavoidable because there might be no full sentences with proper punctuation in an article.

8.3 Language classification approach

As discussed in Chapter 4, we focus on analysis of the last part of sentences. A tool we wrote for this purpose parses the output from MeCab, which contains one token per line (see Chapter 6 for details). The tool starts with the last line and examines as many lines as necessary to classify the sentence-final elements. An example of MeCab output is shown in Table 6.2, on page 53, and the last two lines in the table are always discarded. Looking at the distribution of the different punctuation characters might be interesting, but currently the tool only uses the punctuation character to know when it can start analyzing the preceding lines. The presence of any sentence-final particles is noted, but ignored for the purpose of classifying a sentence. The information on word classes in the MeCab output is then combined with the tokenized morphemes to classify the sentence-final word type. In the case of verbs, the occurrence of *masu* or *mashi* auxiliary verb forms is used to detect whether plain or polite verb forms are used (see Section 6.2.4). A similar approach is used for nouns to determine if polite *desu* or *deshita* copula forms are used, or whether there is ellipsis of the copula. We ignore information about tense.

Upon having classified a sentence, the tool outputs the results, which are stored to disk for later statistical analysis. In addition to classifying the end of the sentence, the tool performs a pass over all lines in the MeCab output. This step was designed to detect elements such as personal pronouns, but is currently incomplete and only used to identify usage of the colloquial first person pronoun *ore*.

Characteristic	Percentage	Cumulative percentage
verb (plain)	60.90	60.90
noun+none	21.60	82.50
dearu	10.77	93.27
iadj+none	2.21	95.49
(unclassified)	1.01	96.51
noun+da	0.64	97.15
shimau	0.24	97.40
noun+dneg (plain)	0.23	97.63
verb+masu	0.12	97.76
dearou	0.06	97.83
nadj+dneg (plain)	0.04	97.87
nadj+da	0.04	97.92
verb+yasui	0.04	97.96
noun+desu	0.03	98.00
verb+te+kudasai	0.02	98.02
noun+darou	0.01	98.04
verb+nikui	0.01	98.05

Table 8.1: Sentence characteristics, encyclopedia articles

8.4 Initial classifier distribution

Applying the classification approach from Chapter 5, we first identify the sentence characteristics, using the tool described above. Some of the most frequently occurring characteristics in sentences in the encyclopedia articles are shown in Table 8.1. The list is a subset of the sentence categories found in the file, and the cumulative percentage only includes the values actually listed. The entry named *(unclassified)* corresponds to sentence patterns not supported by our tool. We added support for the most frequently occurring entries, leaving 1.01% of all sentences in the encyclopedia data unclassified. The characteristics generally name the class of the last word in the sentence and is followed by the name of a Japanese particle or auxiliary verb. For example, sentences classified as *noun+da* end with a noun and the *da* copula form. The *(plain)* value indicates usage of plain verb forms, while *dneg (plain)* corresponds to the negative form *dehanai*.

As many as 60.90% of all sentences in the encyclopedia articles end in

a plain verb. Having a verb in the sentence-final position is consistent with the description of Japanese as a SOV language (see Section 3.2). The *de aru* copula form is also used quite extensively, being the sentence-final element in 10.77% of all sentences, while the *dearou* form occurs in 0.06% of all sentences. Usage of plain verb forms and the *de aru* copula confirms the proper writing style for Wikipedia (see Section 7.1), and what is commonly used in similar writing styles without a specific recipient, such as newspapers and scholarly articles (see Section 3.1.1). There are however 8033 sentences that use polite verbs forms (0.12%), and we study these in more detail below.

For nouns, ellipsis of the copula is most frequent, occurring in 21.60% of the sentences, but 0.64% of sentences have the *da* copula form after a noun. According to the Wikipedia style guideline, the *da* form should be used, but as mentioned in Section 3.1.1, ellipsis of the copula is not uncommon in newspapers or similar writing styles. One possible explanation for this discrepancy is that what is perceived as the correct genre by native speakers, combined with the large amount of text already existing in this style in Wikipedia, has greater influence on the writing style used by editors than the very brief description in the guidelines.

A similar overview for the discussion and user pages is given in Table 8.2, and Table 8.3, respectively. The *te+masu* entry indicates use of the shorter *te-masu* form instead of *te-imasu*. Frequently occurring expressions and interjections are shown between quote marks. These occur often in the user pages, which obviously have the viewpoint of the writer: *yokoso* (welcome), *hajimemashite* (nice to meet you), and *konnichiha* (hello), are the type of expressions that one might expect to find on a personal page.

The writing style of the discussion and user page is not as consistent as the encyclopedia sentences. In the discussion pages, 44.50% of verbs use the polite *masu* forms[2], but 6.24% use plain forms, compared to only 0.12% using the non-dominant form in the encyclopedia data. After nouns, the *desu*

[2]This does not include the *te+masu* verbs, which come in addition.

Characteristic	Percentage	Cumulative percentage
verb+masu	44.50	44.50
noun+desu	12.47	56.97
(unclassified)	6.79	63.76
verb (plain)	6.24	70.01
noun+none	5.52	75.54
noun+desyou	4.69	80.23
verb+desyou	2.68	82.91
verb+te+kudasai	1.95	84.86
nadj+desu	1.61	86.47
iadj+desu	0.92	87.40
noun+dneg+masu	0.84	88.24
dearu	0.82	89.06
noun+kudasai	0.66	89.73
iadj+none	0.65	90.38
noun+da	0.37	90.76
te+masu	0.34	91.10
"arigatou gozaimashita"	0.26	91.37
noun+itashimashita	0.26	91.63
o+verbstem+kudasai	0.22	91.86

Table 8.2: Sentence characteristics, discussion pages

Characteristic	Percentage	Cumulative percentage
verb+masu	48.77	48.77
"yokoso"	6.28	55.06
noun+desu	4.79	59.85
"hajimemashite"	4.30	64.16
verb+te+kudasai	4.23	68.39
noun+none	4.07	72.47
noun+kudasai	3.72	76.19
verb (plain)	3.40	79.60
"konnichiha"	3.29	82.89
(unclassified)	3.28	86.18
nadj+desu	2.05	88.24
o+verbstem+kudasai	1.74	89.99
noun+dneg+masu	0.70	91.71
dearu	0.64	92.35
iadj+desu	0.38	92.73
"arigatou gozaimashita"	0.35	93.08
iadj+none	0.25	93.68
noun+itashimashita	0.19	93.87
noun+da	0.14	94.02

Table 8.3: Sentence characteristics, user pages

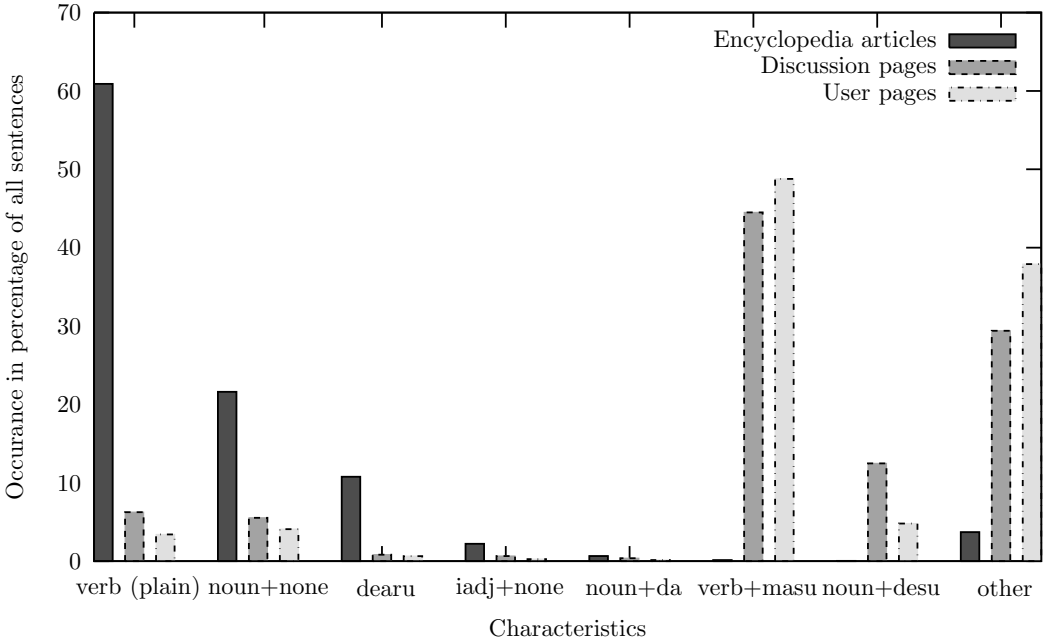


Figure 8.4: Initial classifier distribution

copula form is dominant, at 12.47%, but as many as 5.52% have ellipsis of the copula (the *da* copula form is only used in 0.37% of all sentences). A similar variation is seen in the user page data. Polite verb forms are dominant, at 48.77%, but 3.40% use the informal style. Ellipsis of the copula occurs in 4.07% of all sentences; almost as frequently as the *desu* form, at 4.79%.

The tables show that the writing style of the encyclopedia content is clearly distinct from that of the discussion and user pages. A comparison of the three page types is shown in Figure 8.4, which lists the characteristics that occur most frequently in the encyclopedia articles. Especially the difference between the frequency of plain and polite *-masu* verb forms can be clearly seen. The writing style of the encyclopedia pages is also more consistent, with only a small number of characteristics needed to describe most of the sentences. Interestingly, Emigh and Herring (2005) found that the language in the discussion pages for the English Wikipedia were less formal that of the encyclopedia articles. For the Japanese Wikipedia, we can see that more

Text type	*ga*	*ka*	*ne*	*yo*	*dots*
Wikipedia articles	0.019	0.002	0.001	0.002	0.064
Discussion pages	2.337	0.291	3.324	1.323	0.723
User pages	0.371	0.003	0.851	0.459	0.195

Table 8.4: Sentence-final particles

polite forms are used in the user and discussion pages. However, the difference between the encyclopedia articles and the other two page types is that the first has no specified reader, while the content of the other pages are primarily meant for the other editors. This difference is likely what determines the writing style. The use of plain forms in the encyclopedia articles and polite forms in the other pages does not imply that the level of formality is the opposite of that in the English Wikipedia; the presence of the *te-masu* form is an indication of less formality, despite polite forms being used.

As for sentence-final particle usage, Table 8.4 shows an overview of how frequently the listed particles occur. More than one particle can be used in the same sentence (e.g., *yo* and *ne*), but here the frequency for each particle is given independently of the others. The *dots* field lists usage of more than one punctuation character at the end of a sentence (e.g., "..."). In the encyclopedia articles, particle usage is almost non-existent, especially with regards to the particles *ne* and *yo*. They occur most frequently in the discussion pages, where the particle *ne* occurs in 3.324% of all sentences. These results are consistent with the description in Section 3.1.1, with these particles generally occurring in speech but not written language. All the text here is written, but the discussion pages are closer to spoken conversation than the encyclopedia articles.

Finally, in Table 8.5, we compare some of the alternative characteristics for the encyclopedia articles. For clarity we retain the percentage values from the tables above, but in addition we show the ratio of the less frequently occurring alternatives relative to that of the most frequent. A low ratio value implies a clear difference between how often forms are used. The *not used*

Element	Alternatives (percentages)		
Verb forms,	*plain*	*masu*	
(sentence-final)	60.90%	0.12%	
Ratio	1	0.00197	
Copula forms,	none	*da*	*desu*
(after nouns)	21.60%	0.64%	0.03%
Ratio	1	0.02962	0.00138
Particle *ga*,	not used	used	
(sentence-final)	(99.981%)	0.019%	
Ratio	1	0.00019	

Table 8.5: Alternative characteristic summary, encyclopedia articles

Element	Alternatives (percentages)		
Verb forms,	*masu*	*plain*	
(sentence-final)	44.50%	6.24%	
Ratio	1	0.14022	
Copula forms,	*desu*	none	*da*
(after nouns)	12.47%	5.52%	0.37%
Ratio	1	0.44266	0.02967
Request forms,	*verb+te+kudasai*	*o+verbstem+kudasai*	
(*kudasai*)	1.95%	0.22%	
Ratio	1	0.11282	
Particle *ga*	not used	used	
(sentence-final)	(97.663%)	2.337%	
Ratio	1	0.02392	

Table 8.6: Alternative characteristic summary, discussion pages

Element	Alternatives (percentages)		
Verb forms,	*masu*	*plain*	
(sentence-final)	48.77%	3.40%	
Ratio	1	0.06971	
Copula forms,	*desu*	none	*da*
(after nouns)	4.79%	4.07%	0.14%
Ratio	1	0.84968	0.02922
Request forms,	*verb+te+kudasai*	*o+verbstem+kudasai*	
(*kudasai*)	4.23%	1.74%	
Ratio	1	0.41134	
Particle *ga*,	not used	used	
(sentence-final)	(99.629%)	0.371%	
Ratio	1	0.00372	

Table 8.7: Alternative characteristic summary, user pages

value for the *ga* particle is an estimate.

Compared in this way, the preferred writing style is obvious. Plain verb forms, ellipsis of the copula, and no *ga* particle usage is common in the encyclopedia articles. Note that the table does not include the *de aru* form, where we currently do not keep track of the preceding elements during the analysis. The equivalent values for the discussion pages are shown in Table 8.6, and by comparing the ratio values for the sentence final verb forms, the higher consistency of the encyclopedia data can be seen. The ratio value is 0.00197, compared to 0.14022 for the discussion pages. Two polite request forms are included in this table, with the *verb+te+kudasai* form used more often than the more polite *o+verbstem+kudasai*. The alternatives in the user pages are listed in Table 8.7, where the less polite request form is used quite often (in 4.23% of all sentences), but the *o+verbstem+kudasai* occurs relatively more often than in the discussion pages, having a ratio value of 0.41134, compared to 0.11282 for the discussion pages. By increasing the number of characteristics that we are able to identify, it would be possible to describe genres in more detail, and with a database containing the characteristics of known genres, it might be possible to identify the genre of a text based on the relative frequency of the different alternatives, i.e., using the ratio values.

The results above show that, with regards to the characteristics that we examine, the content of the Wikipedia encyclopedia articles are generally consistent with both what can be expected from written text with no specific recipient, and the Japanese Wikipedia guidelines for proper article writing style. To the extent that there is a discrepancy, it is with the guidelines, that specify usage of the *da* copula form, which occurs to a much lesser degree than ellipsis of the copula. The consistency of the writing style is emphasized by the higher degree of variation in the discussion and user pages, even though the same people produce the content of all three page types. We can conclude that, for the characteristics that we study here, the writing style used in the Wikipedia encyclopedia articles is, with some exceptions,

quite consistent, despite the openness of Wikipedia. If this is the result of some editors correcting style errors, or due to most editors being aware of the proper style, would require further study to determine, but whatever the reason, the results are impressive, considering the way the articles are written.

8.5 Genre violations

The encyclopedia articles mostly use plain forms, but we classified 0.12% sentences as having *masu* forms, and 0.03% as using the *desu* form after a noun. To determine if these sentences represent examples of the guidelines not being followed, we wrote a simple tool that lists the offending sentences and the name of the articles they occur in. We then manually examined the output.

The first observation we made was that there were some deficiencies in the preprocessing stage used to remove unwanted content. Some article types, such as template and image pages, that it would have been correct to remove, were treated as encyclopedia articles. There were also several cases of cited text not having been removed, sometimes due to mismatching or missing parenthesis characters in the text. More problematic for the cleaning approach we use is the observation that in some cases, entire sections contained text written entirely in polite forms. Sections consisting primarily of text from other sources, such as letters, sometimes contain polite forms. Story summaries, e.g., for a book or movie, are also in some cases written entirely in polite forms. Removing this type of text would be difficult, because there is no guarantee that the section titles are named in a consistent way. Additionally, we identified comments regarding the contents of an article having been made by editors, which it most likely would be possible to remove by extending the functionality of the program that performs the preprocessing step.

Finally, there are the actual cases of incorrect writing style. We manually

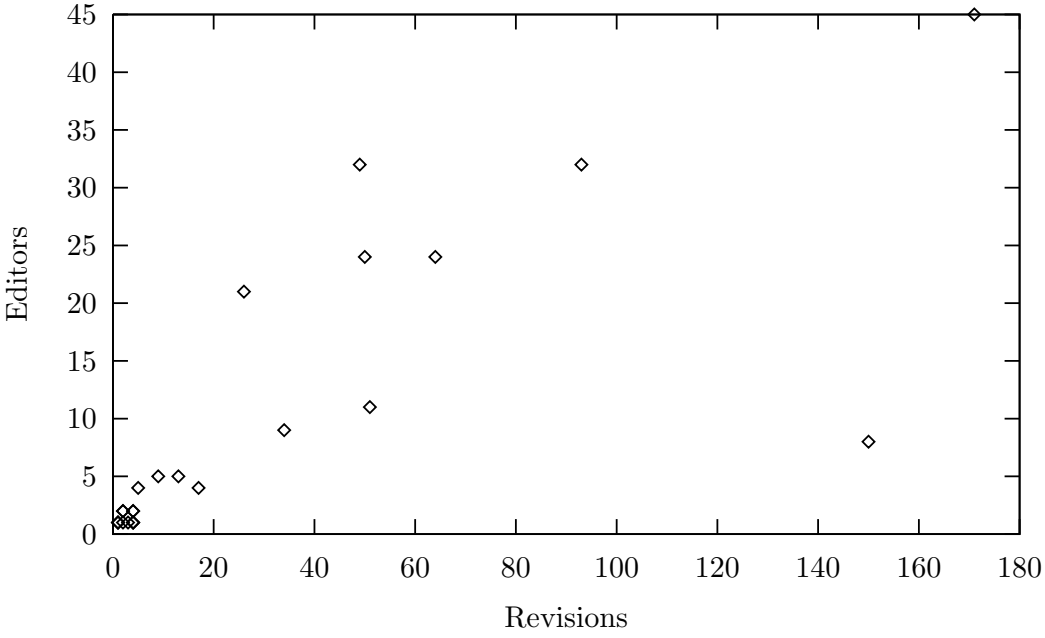


Figure 8.5: Style error distribution (subset)

verified 25 instances of this occurring[3]. We did not examine the entire list of possible style errors, because even after ignoring the obvious instances of sentences that should have been removed during preprocessing, the list contained several thousand entries. The impression given by an examination of a subset of the list is that a large part of the sentences that use polite forms do not represent examples of inconsistent language usage, but text that either should have been removed during preprocessing or text where an editor might argue that polite forms are proper.

The remaining 25 cases, where we determined that usage of a polite form was inconsistent with the surrounding text, are shown in Figure 8.5. Based on the overview of related work, we concluded in Section 2.3 that errors should primarily occur in articles that have few editors and few edits. The figure shows a scatter plot of the number of editors and revisions for each of

[3]In some cases the errors have even been corrected in more recent versions of the pages we identified errors in.

the articles where we identified errors[4], and to a certain extent the results fulfill these expectations. As many as 16 of the 25 errors occur in articles with at most 17 revisions, and no more than 5 editors. As noted above, this is only a subset of the possible errors identified, but the relatively small number of lines using polite forms[5], show that the total number of errors cannot be high.

The difficulty of removing all sentences that correctly make use of polite verb forms would due to many false positives make the approach we have used here less practical for editors that wish to correct these sentences, but the classification procedure we have used correctly determined that while plain forms are most frequent, there are articles where polite forms are used.

8.6 Summary

This chapter presents the results of our examination of language usage in Wikipedia. We examine the encyclopedia pages, the user pages, and the discussion pages, and find that there are clear differences in writing style. The classification procedure described in Section 5.4 is used to describe some aspects of the writing style of Wikipedia, such as use of plain verb forms. We show that there are some cases of errors where polite verb forms are used, but that at least some of these, which we manually inspected, have few edits and few editors. Chapter 2 describes this type of article as most likely to contain errors.

[4]This information was obtained from the file *jawiki-20080724-stub-meta-history.xml*.

[5]After excluding articles and text that it can be easily determined should have been removed during the preprocessing step, there were in total 3550 lines in 1170 articles that used plain forms, which is only 0.054% of all the extracted lines. Only a subset of these lines are actual errors.

Chapter 9

Conclusions

In this Chapter, we present our summary and conclusions, based on the results that we have obtained. We discuss the benefits and drawbacks of the approach that we have used, and list possible areas for future work.

9.1 Summary

The basis for this thesis is the observation that honorifics processes are encoded in a very explicit manner in Japanese. For example, a speaker of Japanese needs to make a conscious choice between plain forms or use of the polite *masu* verbal endings, and the copula has several possible forms, including the plain *da* form and the polite *desu* form. The proper form to use is determined by genre, and our hypothesis has been that it should be possible to describe and identify the genre of a text based on identification of honorifics processes used in a text. Moreover, that the honorifics processes in Japanese are sufficiently explicit for it to be possible to perform the identification process with a computer.

To test this hypothesis we have applied it to Wikipedia, which is a large Internet based encyclopedia that anyone can contribute to. This approach to creating an encyclopedia is very different from the traditional approach,

based on known subject experts writing articles. The openness increases the potential number of contributors, but raises questions as to whether the content can be trusted. Several studies have been made to answer this question, primarily based on having subject experts review a subset of articles. Proper writing style is one aspect of correctness, and assuming our hypothesis is correct, one that can be examined for all articles in the Japanese Wikipedia.

To verify our hypothesis we started by analyzing Japanese honorific processes. The three primary types of honorifics in Japanese are addressee controlled honorifics, subject honorifics, and object honorifics. In addition comes related language features such as beautification, choice of pronominal forms, and sentence-final particles. Addressee controlled honorifics are relatively easy to detect, because they usually control the form of the sentence-ending elements. Use of sentence-final particles can also be determined by studying the end of a sentence and are easy to identify. Pronominal forms require all words in a sentence to be examined, but no additional information. However, the remaining processes are more difficult to detect, especially when ellipsis results in ambiguity. For example, the subject honorific *-rare* suffix is homophonous with the suffix used in passive and potential forms, and which of these is actually used can be impossible to determine unless the context in which a sentence occurs is considered. The difficulty of analyzing sentences with a computer increases with the amount of understanding required to identify honorifics processes reliably. For this reason we chose to focus on the elements that can be identified by studying the final part of a sentence: use of plain or polite forms, and sentence-final particles.

We examined two classification systems described in related work and found these to be too limited to capture the wide range of language usage characteristic that together might be used to provide a detailed description of a genre. Rather than having a fixed system, we chose a classification procedure based on identifying the most frequently occurring characteristics, and then contrasting these with alternatives, when they exist.

For computer based analysis, a challenge with Japanese is the lack of separation between words in a sentence. A morphological analyzer that performs tokenization is typically used to separate a sentence into smaller units that can be used to perform more advanced analysis. We compared the performance of three programs that can tokenize Japanese sentences. A program called MeCab required the least time to analyze a file containing sentences extracted from Wikipedia. According to related work, this tool is also more accurate than at least one of the alternatives, while the third tool suffered from significant performance issues, making the choice of tool simple. Using MeCab, we confirmed the difficulty of correct classification of honorifics processes such as the *o-verb suru* construct. Additional issues are caused by the need for dictionary information during morphological analysis, some of which would require a more in-depth analysis of the dictionary used by MeCab to solve. However, the information provided by MeCab is sufficient for the simple sentence-final analysis we focus on.

Before using MeCab to analyze Wikipedia, we conducted a preliminary analysis. We identified what Wikipedia guidelines give as the proper writing style, being plain verb forms, and *de aru* and *da* copula forms. Through a study of the content of the downloadable XML files that contain snapshots of all Wikipedia articles we identified elements such as markup and lists that must be removed before the language in an article can be classified.

Finally, we applied MeCab to the text extracted from a Wikipedia snapshot file from July 24, 2008, and analyzed the results with a tool we wrote for this purpose. The results show that language usage in Wikipedia is fairly consistent and follows the official guidelines. The biggest discrepancy is with regards to the *da* copula form, which is listed in the guidelines, but occurs less frequently than ellipsis of the copula. A comparison of the encyclopedia text with the content of the Wikipedia discussion and user pages shows that the style of the encyclopedia articles is distinct from the other two, and applied much more consistently. Only a small number of sentences in the

encyclopedia articles make use of polite verb and copula forms, and manual examination of these sentences revealed that they were to a large extent caused by insufficient preprocessing of the content. While we did not examine all the lines with polite forms, we identified 25 instances where we determined that polite forms were likely to have been used incorrectly. For these instances, the majority were articles that had only had a small number of editors and revisions. That errors are likely to exist in this type of article is consistent with our expectations based on study of related work.

9.2 Evaluation of thesis claims

In Chapter 1, we made four claims. We now examine the extent to which these can be said to have been validated.

Firstly, that the presence or absence of honorific processes can be used to classify the genre of a text. For addressee honorifics, we consider this to be the case. Especially for plain and polite forms, a choice needs to be made with regards to which form to use. Our comparison of the different Wikipedia page types also shows that there are differences between usage of these forms, which can be considered one aspect of genre. For subject and object honorifics, identification is more problematic. As discussed in Chapter 4, it is difficult to detect the non-usage of these types of honorifics in a context where usage would be proper, especially in cases of ellipsis. For a native speaker these cases might be obvious, but even in the cases where this can be determined from the surrounding text it would not be trivial for a computer program to do so. Additionally, there are potential difficulties with identifying the constructs used by some of the honorifics processes due to ambiguity. Resolving this cases might be possible but we have not attempted to do so in this thesis. The results of these limitations is that some characteristics relevant to genre cannot easily be identified.

For our analysis of Wikipedia, these limitations are less problematic be-

cause the text does not have a known recipient, but is written with no specific reader in mind. However, for comprehensive analysis of personal letters or spoken language, usage of subject or object honorifics would be more important, even though a study to determine whether these processes are actually used in Wikipedia might be interesting.

Our second claim is that the analysis can be automated and performed with a computer. Chapter 8 provides the result of this type of analysis, and shows that even though we had to limit the characteristics that we studied, it is possible to analyze even millions of sentences and obtain useful results.

The same chapter also shows that the third claim is possible; describing the writing style used in a project like Wikipedia, and the extent to which it is applied consistently. Our results show clear differences between the three Wikipedia page types, and that the writing style of the encyclopedia articles is largely consistent with what is specified in the Wikipedia guidelines.

The fourth claim concerns the possibility of identifying incorrect language usage. In Chapter 8, we present examples of this, but the difficulty of correctly extracting only the encyclopedia text gave many false positives. The preprocessing step can likely be improved, but the way in which some sections are written in polite forms, without this necessarily being inconsistent with the style guidelines, makes it difficult to ensure that false positives do not occur, at least in the case of Wikipedia. For text data from other sources this problem might not apply.

9.3 Conclusions

In this thesis, we have studied automated language characterization and the correctness of Wikipedia articles. We have found that there are clearly aspects of language usage that it is difficult to identify with a computer, but even characteristics that are easy to identify can be useful in examining the content of a large project such as Wikipedia. Despite being a distributed

encyclopedia with a large number of contributors, we were able to identify only a small number of what we determined to be actual errors in writing style. This was partly due to false positives making it necessary to manually confirm the errors, but even including the false positives, the low number of errors is still impressive.

9.4 Future work

Several problems and limitations were encountered during the work on this thesis, providing ample room for future work on all steps in the analysis procedure, including the following:

- Especially in the case of Wikipedia, improving the content extraction procedure would increase the correctness of the classification and detection of inconsistent language usage. Related to this would be a continuation of the study we have done here on the Wikipedia data. We have determined that incorrect language usage exists, but not studied how these errors are introduced, for how long they exist, and whether they are corrected by a small number of editors, or a large number of anonymous users that only make few corrections. The first would indicate that some users focus on correcting this type of problem, while the second might indicate that readers correct errors that they discover.

- Ambiguity makes several of the honorifics processes, such as the honorific *-rare* prefix, difficult to identify. A more advanced sentence analysis than the one we have applied here might be able to resolve these problems. In general, increasing the number of characteristics that are supported by the classification script would increase the scope for genre identification and description.

- The analysis procedure we have used can be applied to other text types, such as newspaper articles. Using corpus data would additionally elim-

inate the need for the content extraction step, and improve the correctness of the results. Through examination of many different text types, it might be possible to build a database of characteristics found in different genres. This information might again be usable for identifying the genre of an unknown text.

Bibliography

Adler, B. T. and de Alfaro, L. 2007. A content-driven reputation system for the wikipedia. In *WWW '07: Proceedings of the 16th International World Wide Web Conference*, pages 261–270.

Altmann, U. 2005. Representation of medical informatics in the wikipedia and its perspectives. In *MIE '05: Proceedings of the 19th International Congress of the European Federation for Medical Informatics, Connecting Medical Informatics and Bio-Informatics*, pages 755–760.

Anthony, D., Smith, S. W., and Williamson, T. 2007. The quality of open source production: Zealots and good samaritans in the case of wikipedia. Technical Report TR2007-606, Dartmouth College, Computer Science.

Asahara, M. and Matsumoto, Y. 2004. Japanese unknown word identification by character-based chunking. In *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, pages 459–465.

Blumenstock, J. E. 2008. Size matters: word count as a measure of quality on wikipedia. In *WWW '08: Proceedings of the 17th International Conference on World Wide Web*, pages 1095–1096.

Bray, T., Paoli, J., Sperberg-McQueen, C. M., Maler, E., and Yergeau, F. 2008. Extensible markup language (xml) 1.0 (fifth edition). Technical report, The World Wide Web Consortium.

Chesney, T. 2007. An empirical examination of wikipedia's credibility. *First Monday*, 11(11).

Clauson, K. A., Polen, H. H., Boulos, M. N. K., and Dzenowagis, J. H. 2008. Scope, completeness, and accuracy of drug information in wikipedia. *The Annals of Pharmacotherapy*, 42(12):1814–1821.

Cook, H. M. 1998. Situational meanings of japanese social deixis: The mixed use of the masu and plain forms. *Journal of Linguistic Anthropology*, 8(1):87–110.

Den, Y., Nakamura, J., Ogiso, T., and Ogura, H. 2008. A proper approach to japanese morphological analysis: Dictionary, model, and evaluation. In *LREC '08: Proceedings of the Sixth International Language Resources and Evaluation*.

Denning, P. J., Horning, J., Parnas, D. L., and Weinstein, L. 2005. Wikipedia risks. *Communications of the ACM*, 48(12):152.

Devgan, L., Powe, N., Blakey, B., and Makary, M. 2007. Wiki-surgery? internal validity of wikipedia as a medical and surgical reference. *Journal of the American College of Surgeons*, 205(3).

Dondio, P. and Barrett, S. 2007. Computational trust in web content quality: A comparative evalutation on the wikipedia project. *Informatica*, 31:151–160.

Dridan, R. and Baldwin, T. 2007. What to classify and how: Experiments in question classification for japanese. In *PACLING '07: Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, pages 333–41.

Emigh, W. G. and Herring, S. C. 2005. Collaborative authoring on the web: A genre analysis of online encyclopedias. In *HICSS-38 '05: Proceedings of the 38th Hawaii International Conference on System Sciences*.

Fuchi, T. and Takagi, S. 1998. Japanese morphological analyzer using word co-occurrence: Jtag. In *COLING '98: Proceedings of the 17th international conference on Computational linguistics*, pages 409–413.

Giles, J. 2005. Internet encyclopaedias go head to head. *Nature*, 438:900–901.

HU, M., Lim, E.-P., Sun, A., Lauw, H. W., and Vuong, B.-Q. 2007. Measuring article quality in wikipedia: Models and evaluation. In *CIKM '07: Proceedings of the sixteenth ACM Conference on information and knowledge management*.

Imamura, K., Kikui, G., and Yasuda, N. 2007. Japanese dependency parsing using sequential labeling for semi-spoken language. In *ACL '07: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 225–228.

Ivana, A. and Sakai, H. 2007. Honorification and light verbs in japanese. *Journal of East Asian Linguistics*, 16(3):171–191.

Kacmarcik, G., Brockett, C., and Suzuki, H. 2000. Robust segmentation of japanese text into a lattice for parsing. In *COLING '00: Proceedings of the 18th conference on Computational linguistics*, pages 390–396.

Kaiser, S., Ichikawa, Y., Kobayashi, N., and Yamamoto, H. 2001. *Japanese: A Comprehensive Grammar*. Routledge.

Kittur, A., Chi, E., Pendleton, B. A., Suh, B., and Mytkowicz, T. 2007. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. In *CHI '07: Proceedings of the 25th Annual ACM Conference on Human Factors in Computing Systems*.

Kudo, T. and Matsumoto, Y. 2002. Japanese dependency analysis using cascaded chunking. In *CoNLL '02: Proceedings of the 6th Conference on Natural Language Learning*, pages 63–69.

Kudo, T., Yamamoto, K., and Matsumoto, Y. 2004. Applying conditional random fields to japanese morphological analysis. In Lin, D. and Wu, D., editors, *EMNLP '04: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237.

Lih, A. 2004. Wikipedia as participatory journalism: Reliable sources? metrics for evaluating collaborative media as a news resource. In *Proceedings of the 5th International Symposium on Online Journalism.*

Lorenzen, M. 2006. Vandals, administrators, and sockpuppets, oh my! an ethnographic study of wikipedia's handling of problem behavior. *MLA Forum*, 5(2).

Lucas, N. 1991. *European studies in Japanese linguistics, 1988-90*, chapter The suffix verb DEARU in scientific and technical literature, pages 63–74. Lone Publications.

Luyt, B., Aaron, T. C. H., Thian, L. H., and Hong, C. K. 2008. Improving wikipedia's accuracy: Is edit age a solution? *Journal of the American Society for Information Science and Technology*, 59(2):318–330.

Luyt, B., Kwek, W. T., Sim, J. W., and York, P. 2007. Evaluating the comprehensiveness of wikipedia: The case of biochemistry. In *Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers*, volume 4822, pages 512–513.

Maeda, H., Kato, S., Kogure, K., and Iida, H. 1988. Parsing japanese honorifics in unification-based grammar. In *ACL '88: Proceedings of the 26th annual meeting on Association for Computational Linguistics*, pages 139–146.

Makino, S. and Tsutsui, M. 2002a. *A Dictionary of Basic Japanese Grammar.* The Japan Times, Ltd.

Makino, S. and Tsutsui, M. 2002b. *A Dictionary of Intermediate Japanese Grammar*. The Japan Times, Ltd.

Makino, S. and Tsutsui, M. 2008. *A Dictionary of Advanced Japanese Grammar*. The Japan Times.

Matthews, P. H. 2007. *The Concise Oxford Dictionary of Linguistics*. Oxford University Press, USA.

Mayumi, U. 2002. *Discourse Politeness in Japanese Conversation: Some implications for a Universal Theory of Politeness*. Hituzi Syobo Publishing LTD.

Mcgloin, N. H. 1990. *Aspects of Japanese Women's Language*, chapter Sex Difference and Sentence-Final Particles. Kurosio Publishing.

McGuinness, D. L., Zeng, H., da Silva, P. P., Ding, L., Narayanan, D., and Bhaowal, M. 2006. Investigations into trust for collaborative information repositories: A wikipedia case study. In *MTW '06: Proceedings of the WWW '06 Workshop on Models of Trust for the Web*.

Murata, M., Uchimoto, K., Ma, Q., and Isahara, H. 2000. Bunsetsu identification using category-exclusive rules. In *COLING '00: Proceedings of the 18th conference on Computational linguistics*, pages 565–571.

Musteric, M. 2003. *Japanese/Korean Linguistics*, volume 11, chapter Honorifics in a Democratized Japan: Changes in the Usage of Keigo Since World War II (A Case Study of Letters and Literature), pages 163–174. CSLI Publications.

Nielsen, F. A. 2007. Scientific citations in wikipedia. *First Monday*, 12(8).

Ortega, F., Gonzalez-Barahona, J. M., and Robles, G. 2008. On the inequality of contributions to wikipedia. In *HICSS '08: Proceedings of the 41st Annual Hawaii International Conference on System Sciences*.

Oshima, D. Y. 2008. *Japanese/Korean Linguistics*, volume 13, chapter Semantic Divergence of -(R)are: From a Different Perspective, pages 309–320. CSLI Publications.

Potthast, M., Stein, B., and Gerling, R. 2008. Automatic vandalism detection in wikipedia. In *ECIR '08: Proceedings of the 30th European Conference on IR Research, Advances in Information Retrieval*, pages 663–668.

Rector, L. H. 2008. Comparison of wikipedia and other encyclopedias for accuracy, breadth, and depth in historical articles. *Reference Services Review*, 36(1):7–22.

Sanger, L. 2005. *Open Sources 2.0: The Continuing Evolution*, chapter The Early History of Nupedia and Wikipedia: A Memoir. O'Reilly.

Sasano, R. and Kurohashi, S. 2008. Japanese named entity recognition using structural natural language processing. In *IJCNLP '08: Proceedings of the 3th International Joint Conference on Natural Language Processing*, pages 607–612.

Shibatani, M. 1991. *The Languages of Japan*. Cambridge Language Surveys.

Shirado, T., Marumoto, S., Murata, M., Uchimoto, K., and Isahara, H. 2006. A system to indicate honorific misuse in spoken japanese. In *ICCPOL '06: Proceedings of the 21st International Conference of Computer Processing of Oriental Languages. Beyond the Orient: The Research Challenges Ahead*, volume 4285, pages 403–413.

Stvilia, B., Twidale, M. B., Smith, L. C., and Gasser, L. 2008. Information quality work organization in wikipedia. *Journal of the American Society for Information Science and Technology (JASIST)*, 59(6):983–1001.

Tamura, A., Takamura, H., and Okumura, M. 2007. Japanese dependency analysis using the ancestor-descendant relation. In *EMNLP-CoNLL '07:*

Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 600–609.

Tatematsu, K., Tateoka, Y., Matsumoto, T., and Sato, T. 1997. *Writing Letters in Japanese*. The Japan Times, Ltd.

Utsuro, T., Nishiokayama, S., Fujio, M., and Matsumoto, Y. 2000. Analyzing dependencies of japanese subordinate clauses based on statistics of scope embedding preference. In *NAACL '00: Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 110–117.

Viégas, F. B., Wattenberg, M., and Dave, K. 2004. Studying cooperation and conflict between authors with history flow visualizations. In *CHI '04: Proceedings of the 2004 Conference on Human Factors in Computing Systems*, pages 575–582.

Viégas, F. B., Wattenberg, M., Kriss, J., and van Ham, F. 2007a. Talk before you type: Coordination in wikipedia. In *HICSS-40 '07: Proceedings of the 40th Hawaii International International Conference on Systems Science*, pages 78–88.

Viégas, F. B., Wattenberg, M., and McKeon, M. M. 2007b. The hidden order of wikipedia. In *OCSC '07: Proceedings of Online Communities and Social Computing, Second International Conference*, pages 445–454.

Voss, J. 2005. Measuring wikipedia. In *ISSI '05: Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics*.

Waters, N. L. 2007. Why you can't cite wikipedia in my class. *Communications of the ACM*, 50(9):15–17.

Wetzel, P. 2004. *Keigo in Modern Japan.* University of Hawaii Press.

Wilkinson, D. M. and Huberman, B. A. 2007. Cooperation and quality in wikipedia. In *WikiSym '07: Proceedings of the 2007 international symposium on Wikis*, pages 157–164.

Yoshimoto, K. 1988. Identifying zero pronouns in japanese dialogue. In *COLING '88: Proceedings of the 12th conference on Computational linguistics*, pages 779–784.

Sammendrag

Det Internet-baserte leksikonet Wikipedia er en potensielt veldig nyttig informasjonskilde, men det er naturlig å være skeptisk til et leksikon der alle kan gjøre endringer. Flere undersøkelser har vært gjort for å finne ut hvor korrekt innholdet i Wikipedia er, men det store antallet artikler og de kontinuerlige endringene begrenser hvor omfattende denne typen undersøkelse kan være.

For et leksikon er skrivestil et aspekt ved korrekthet, og spesielt for Wikipedia vil en uformell eller lite konsistent skrivestil gi et dårlig inntrykk. Hvis feil som kan rettes av enhver som leser teksten ikke blir fjernet, hvor sannsynlig er det at feil blir rettet i informasjon som bare en ekspert på emnet kan oppdage?

Vi har undersøkt den japanske Wikipedia, fordi japansk er et språk der indikatorer på høflig og formell tale forekommer veldig eksplisitt, via forskjellige former som det i noen tilfeller er nødvendig å velge mellom når man snakker eller skriver. Spesielt er forskjellen mellom vanlige og høflige former enkle nok å identifisere til at at det kan gjøres med en datamaskin, noe som muliggjør en undersøkelse av all teksten i alle de japanske Wikipedia artiklene.

Ved hjelp av denne framgangsmåten fant vi ut at skrivestilen i Wikipedia er stort sett i tråd med retningslinjene for prosjektet. De unntakene vi undersøkte forekom hovedsakelig i artikler som bare hadde hatt et lite antall endringer og få forskjellige forfattere.